目 录

第一章 绪 论

§1 数学地质的产生及其现代含义

地质学是一门以地壳为研究对象，有着悠久历史的自然科学。但是，与其它学科相比，它的定量化程度直到目前仍然是比较低的，造成这种状况的根本原因是地质因素的多样性、地质过程的不可再现性以及它所遗留下来的地质信息的片面性。也就是说，地质学的研究内容，几乎都是发生在地球历史中的一些地质过程，而目前可以观察或测量的一些地质现象都是经历了长期的地质演化过程以后，所遗留下来的地质过程的残留记录，人们只能根据这些残留记录提供的部分地质信息去推断早已发生的地质过程。因此，有相当长的一个时期，地质学的研究方法是：首先对地质现象进行观察、记录和描述，收集实际的地质资料，然后再进行分类归纳和逻辑推理，最后得出相应的地质认识或地质论断。目前的多数地质理论就是由上述研究方法得出、而后又经大量事实所证实了的具有规律性的地质论断。

对于同一个地质问题，应用传统的定性研究方法，会使人们收集的地质资料有所差异，甚至是在同一批地质资料的条件下，也会因地质学家思维方法的差异而得出不同的结论；另外，这种传统的定性研究方法，也就决定了地质学在相当长的一个时期内只能是一门定性的科学。为改变地质学的上述状况，以适应生产发展的需要，必然促使一些人把地质学和其它基础自然科学相结合，以期对所研究的地质问题作出合理的地质解释。本世纪以来，地质学与物理学、化学以及力学相结合，结果产生了地球物理学、地球化学和地质力学以及相应的地质勘探方法。这些新的边缘学科的产生和发展，极大地促进了地质学的发展，使古老的地质学产生了质的飞跃，表现出强大的生命力。

由于测试仪器的不断改进和更新，新的边缘学科所产生的新的物、化探手段不断增加，从而可以观察到从宏观到微观的一切地质的、物理的和化学的特征，使不同类型的地质信息巨增，特别是出现了大量的数值型资料，这就使地质人员无法应用定性的研究方法去处理、分析和利用这些资料，甚至连对资料的阅读都成为十分困难的事，因而大量有用的地质信息，特别是那些数值型的信息就被浪费掉。为了充分利用地质勘探中所获得的各种找矿信息，加强地质研究，减少勘探风险，这就必须在地质学中引人定量的研究方法。地质学中应用数学方法大约始于19世纪40年代初期，但定量化的进程却非常缓慢。

欲想得到地质问题的定量结论，这不仅需要引人定量的数学方法，而且还应有记忆能力强、处理速度快的资料处理设备。50年代初，计算机的批量生产和数字绘图仪的问世，为在地质学中引人定量研究方法创造了条件。找矿难度的不断加大、不同类型地质信息的巨增以及计算机技术的普及，又进一步促进了地质学与数学的结合。从本世纪50年代末期开始，数学方法和电子计算机在地质学中获得广泛应用，到60年代末期形成一门新的边缘学科——数学地质。多数的数学地质工作者认为：数学地质是地质学与数学和计算机科学相互渗透、紧密结合而逐步形成的一门地质学的边缘学科。它是以数学为方法，以计算机为主要的研究手段，定量研究地质学基础理论和定量探矿法的一门方法性科学。随着其它科学的发展和生产的需要，数

学地质的理论和方法也在不断地发展和完善，它已从理论研究向应用的方向发展，并已能直接为国民经济服务，其最终目的是实现地质学的定量化和智能化。

石油数学地质是数学地质在石油勘探、开发及石油资源评价等领域中的应用，其主要任务是研究与石油地质学有关的一些问题的定量化，近年来它的主要研究内容是石油资源评价方法及其相应的软件。

数学地质的基本工作过程可以概括为：由地质学家提出地质问题，分析问题的地质因素，建立相应的地质概念模型；选择合适的数学方法，将定性的地质概念模型转化为定量的数学模型并研制相应的应用软件；对计算机输出的定量结果及地质图形资料进行地质解释，并在此基础上确定或修改给出的地质概念模型及相应的数学模型，以期解决所提出的地质问题。上述工作流程如图 1-1 所示。

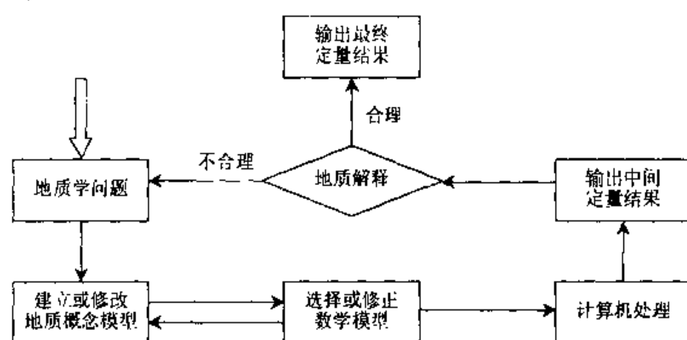


图 1-1　数学地质基本工作流程

§2　数学地质的主要研究内容

一、地质多元统计

地质多元统计是应用宏观统计方法研究地质问题的方法的统称，其中的大多数方法是从数理统计中直接移植来的，少数方法是根据地质工作的实际需要，在移植的基础上逐步发展衍生出来的。

地质多元统计是数学地质的基础，也是石油数学地质的主要方法。本世纪 70 年代以前，数学地质的主要研究内容就是地质多元统计方法及其在地质工作中的应用。

目前，地质多元统计方法已比较完善，它对地质学的定量研究起了极大的促进作用。但是，它在地质学领域中应用的广度和深度还远远不够，而且多数常规地质人员对地质多元统计方法也并不太熟悉。

目前常用的地质多元统计方法有回归分析、趋势面分析、聚类分析、判别分析、因子分析、对应分析、相关分析、时间序列分析、线性映射和马尔可夫概型分析等。

在近年的全国性数学地质学术讨论会上，地质多元统计方面的学术论文数量最多，一般占论文总数的 50%～60%，这就表明地质多元统计仍然是数学地质的重要组成部分，对地质研究工作还在起着重要的作用。地质多元统计在长时间内能够存在并不断发展，这是由地质学本身的特点决定的。因为任何一个地质问题都是非常复杂的，即地质问题具有时间久、空间广和因素多这三个基本特征。因而，地质人员总试图借用统计分析方法从已知信息中获得一些规律性的认识，以便从量的角度研究和分析地质问题，这就决定了地质多元统计在地质研究工作中

具有广阔的应用领域。

二、矿产资源预测

随着社会对矿产资源需求量的不断增加和迅速开采，使找矿的难度和风险也越来越大。从本世纪70年代开始，矿产资源的定量预测就已成为数学地质的重要研究内容之一。实践证明，数学地质定量预测方法的有效性越来越明显，并已得到地质界的普遍重视。

对一个探区进行矿产资源预测，要解决两个基本问题：一是有无矿产？有多少？二是如果有可供开采的矿产，到哪里找？此外，对探区进行经济评价与勘探决策也应作为地质评价的延伸性工作。

矿产资源预测在石油部门称为石油资源评价，其目前的主要内容包括：

(1) 探区石油资源量的估算；

(2) 确定探区中的有利勘探地区；

(3) 石油勘探的经济分析。

目前，我国有关地质找矿部门根据各自的实际需要，针对所找矿产的地质特征，都在大力研究矿产资源的定量预测理论及其相应的计算方法和应用软件。例如，石油部门近年来大力开展盆地数值模拟的理论和方法研究，并已在油气资源评价中起了重要的作用；冶金部门近十年来一直大力发展"地质统计学"。地质统计学是南非金矿矿山工程师克里格(D. G. krige)首先提出的预测金矿矿床的一种方法，后来经过法国的应用数学家马特隆(G. Matheron)加以理论化，而成为一个比较完整、自成体系的研究固体矿床，预测品位空间分布的专门方法。近年来，已在法国及法语系国家形成一支有影响的学派。由于这种方法是由克里格首先提出的，故称其为克里格法。地质统计学的主要研究内容是区域化变量，它是通过变异函数，研究固体矿床的空间变化、勘探方法与储量误差三者之间的数量关系。

三、地质数据库

地质数据库是计算机技术为地质人员服务的一个范例，也是一个国家地质现代化的一个重要标志。地质数据库的出现，使地质人员从繁琐的收集和整理地质资料等非研究性的工作中解脱出来，把精力集中到研究中去。它已成为科学管理地质信息的重要手段，从而为地质问题的定量研究和地质资料处理自动化创造了有利的条件。

数据库是分门别类的存储在一起的相关数据的集合。数据的存储独立于使用它的程序，对于插入新数据，修改和检索原有数据均能按照公用的和可控的方式进行。因而，一个完善的数据库应包括数据的存储、检索、更新、处理、显示、通讯及网络等多种功能。

数据库是本世纪60年代末出现的最新数据管理技术，比较完善的数据库软件系统是70年代初完成的。地质数据库在美国、加拿大、法国、德国等西方国家发展很快，80年代以来，在许多国家已普及应用。目前，世界上大约已建成500多个大、中型地质数据库，这些数据库已涉及到地质学的各个分支领域，其中某些大型数据库已在一个国家甚至许多国家形成网络系统。比较著名的地质数据库有：

1. 计算机化矿产资源信息库(CRIB)

它是美国地质调查所的矿产资源数据库。库内存储了美国的4万多个矿床和矿产产地以及其他国家的6千多个矿床和矿产产地的资料。数据文件中包括：矿床位置、地质特征、储量、产量等多种数据。用户可以通过计算机网络系统在全世界500多个城市用电话查询和索取这个数据库中的地质数据。

2. 北美石油数据系统(PDS)

该数据库包括了目前公开投人使用的10个石油地质数据库。库中存储了美国和加拿大的10万余个油气田的有关资料。文件内容包括:油气田的产量、生产井数、储量、圈闭类型、储层时代、储层厚度、油气性质、地层温度、地层压力、岩性等多种数据。

3. 井史控制系统(WHCS)

该数据库属于美国的“石油信息公司”。它是世界上目前最大的一个油井数据库,存储了美国100多万口油气井的有关数据。

虽然地质数据库的数据结构、层次的命名术语等方面各有不同,但是,数据库目前大的层次基本上可以归纳为4个级别,即子库、文件、项目、数据。

近年来,国内利用微型计算机管理地质数据的工作进展很快。但是,因为微型计算机的功能与存储空间都很有限,所以建立的数据库大体上相当于上述4个级别中的文件。因此,建立在微型计算机基础上的数据库,一般只适用于某些专项地质工作。

四、地质过程的数学模拟

应用数学模拟的方法研究地质历史的演化过程,是探索地质基础理论的重要途径之一。近十年来,地质过程的数学模拟已发展成为数学地质的重要组成部分,其发展速度较快,例如,通过盆地模拟研究石油地质演化史已成为一个热门课题,并直接服务于油气资源评价。在国际数学地质协会第25周年大会上,许多国家的多篇论文都涉及到地质过程的数学模拟这个问题。例如,瑞典 Fookes 介绍了模拟沉积旋回性的程序 EFPRO-EUSTASY,该程序综合考虑了水动型海平面变化,沉积物供给、堆积速率,以及构造沉降、抬升这三种因素,已成为商业程序由一家美国公司出售。捷克地质调查所用PDI(IES)盆地模拟系统研究了西北维也纳盆地外卡帕推覆体的热演化史,其中仔细地估计了逆冲过程中岩石的物性参数及压实作用的变化。这一研究表明,盆地模拟已由张性盆地进入到压性盆地。俄罗斯 Svalova 通过建立在流变性和热力学方程基础上的力学—数学模型推导了地幔底辟与沉积盆地形成的关系,指出从较深处快速上升的地幔底辟将导致地壳浅部沉积盆地的形成,否则会导致地壳浅部上隆,这一结果对地幔底辟将导致沉积盆地形成的传统概念是一个挑战。意大利 AGIP 石油公司介绍了以地质统计学的方法模拟油田的三维岩相分布及各岩相内部孔隙度和渗透率的频率分布,进而开展采油史动力学模拟,以此指导油田开发方案的设计。美国斯坦福大学的哈博教授及他的学生们自60年代以来悉心研究沉积盆地形成过程的模拟。他们研制的 SEDSIM 系统在模拟沉积物剥蚀、搬运及沉积的动力学过程时考虑了河流的流量、波浪作用、沉积物搬运的力学机制,盆地环流、压实、粒间水的挤出以及均衡补偿等作用。

由上述地质过程模拟的例子可知,地质过程数学模拟的关键是确定定性表征地质过程的地质概念模型和定量描述地质概念模型的数学模型。地质概念模型是指在对地质体系深刻理解和抽象思维的基础上,以定性方式表达地质体系发生和演化过程及其量间关系的模型,而地质数学模型是指用定量方法描述地质体系发生、演化过程及其量间关系的模型。地质概念模型是建立数学模型的基础,把概念模型转化为数学模型是对地质体系认识的深化和概括。为了分析概念模型与数学模型的可靠性,经常采用试验方法对模型进行验证。例如,用水槽试验模拟沉积过程;用泥巴试验、光弹试验模拟构造演化过程等。

数学模型按其使用的数学方法又可分为确定型与随机型模型。然而,任何一个地质过程都不可能是单一的确定型过程或随机过程,而是两种地质过程在时间和空间上的叠加。因而,一个完善的数学模型应该是由上述两种模型构成的复合模型。目前,人们侧重于用单一的确定型

或随机型模型来研究地质历史演化过程，对地质体系认识上的不完全或者把复杂的地质体系进行简化描述是造成使用单一模型的根本原因。

地质过程的数学模拟，实际上就是在对地质体系分析、归纳和逻辑思维的基础上，先提出一个表征地质体系的地质概念模型，并把它转化为一个数学模型，然后通过计算机对数学模型进行反复运算，以再现地质过程的发生和演化史，进而实现对地质过程定量描述所做的一种试验。

五、地质绘图自动化

地质绘图占据了地质科技人员大量的时间，使他们不能有更多的时间集中精力综合研究所进行的地质科研课题，因此地质绘图自动化一直是地质科技人员的一个追求目标，也是数学地质的主要研究内容之一。

随着计算机技术和图形学的发展，计算机自动绘制地质图的工作进展很快，已有形成一门独立学科的明显趋势。对于地质工作中的绝大部分图件，目前计算机不仅可以绘制，而且能够绘制出精美的彩色图件。可以说，在不久的将来就会实现地质绘图的自动化。

上述五个方面就是现阶段数学地质的主要研究内容。这五个方面既相互独立又相互联系。虽然每个方面的研究内容各有侧重，但它们都是为了一个统一的目的，那就是加快地质学的定量化进程，最终实现地质学研究的定量化和智能化。

一个学科的定量化程度表征着它的成熟和完善程度。当今科学技术正处在飞速发展的时代，任何一个学科都在汲取数学的成熟方法以及最新成就，或者依据本学科的需要向数学界提出新的问题，从而促进数学的发展并服务于本学科。地质学的定量化就是用数学的语言描述地质学中的定义、概念和规律等，从而使地质学由定性描述转变到全面的定量描述。

实现地质学的定量化是十分困难的。这除了地质学自身的复杂性外，还存在着其他一些难点，例如同一地质概念的多种含义问题，观测手段的精度问题，地质数据的代表性问题等都在很大程度上阻碍着地质学实现定量化。因此，地质学的定量化将需要一个相当长的历史阶段，需要几代人的不懈努力方能逐步实现。

第二章　地质数据与地质变量

§1　地质数据

一、地质数据的概念

地质数据是表示地质信息的数、字母和符号的集合。它是用来表示地质客观事实这一地质信息的。从广义角度来看，地质数据可以是定量、定性数据，也可以是文字说明，甚至是地质图形。从狭义角度来看，地质数据主要是指定量的和定性的地质数据。

二、地质数据的类型

地质数据按其特点可以分为观测数据、综合数据和经验数据三大类。

（一）观测数据

观测数据是指利用各种观测手段对研究对象进行观测或度量所获得的数据。是地质数据的主要类型。这类数据一般未进行任何加工处理，所以也称为原始数据。观测数据根据其本身的特点可分为定量数据和定性数据。

1. 定性数据

定性数据是指不能用数值描述，只能用符号或代码描述的观测数据。这种数据不具备数量上的概念，它包括名义型数据和有序型数据两类。

(1) 名义型数据

名义型数据没有数量上的概念，并且数据之间也没有次序关系，只能用符号或代码形式表示。名义型数据是通过区分不同的对象或个体并赋予不同的代码后形成的。例如描述岩石颜色的红、绿、灰、黑，可用符号 A、B、C、D 表示，又如，若不深究岩性的详细概念，砂岩、泥岩、灰岩等岩石类型可用符号 S、M、L 表示等等。符号 A、B、C、D 和 S、M、L 就是名义型数据。

名义型数据量之间只存在“相等”或“不相等”的关系，如红色等于红色（$A=A$），砂岩不等于灰岩（$S\neq L$）。

(2) 有序型数据

有序型数据虽没有数量上的概念，但数据之间有次序关系，常以等级符号或代码形式表示。例如干酪根的范氏分类法将干酪根分为Ⅰ、Ⅱ、Ⅲ三个级别，分别用符号 1、2、3 表示，又如鉴定岩石相对硬度的摩氏标准，将硬度由小到大分为十个级别，即：滑石、石膏、方解石、萤石、磷灰石、长石、石英、黄玉、刚玉、金刚石，分别用符号 1、2、3、4、5、6、7、8、9、10 表示。

有序型数据之间除有相等、不相等关系外，还有“大于”或“小于”关系。如滑石硬度小于石英硬度，从生油潜力看，Ⅰ型干酪根大于Ⅲ型干酪根。

2. 定量数据

定量数据是指能用数值大小来描述的观测数据。包括间隔型数据和比例型数据两类。

(1) 间隔型数据

间隔型数据有明确的数量概念，可以用数值形式表示。例如以基准海平面起算的地层分层数据就是典型的间隔型数据。

间隔型数据之间除了具有相等、不相等以及大于、小于关系外，还可以定量说明数据之间的差异，这种差异具有实际意义。如某地层底界和顶界分层深度之差等于该地层厚度。

(2) 比例型数据

比例型数据也有明确的数量概念，可以用数值形式表示。比例数据之间不仅其差值具有实际意义，而且比值也有实际意义。它和间隔型数据的另一个区别是比例型数据是以0为边界的定量数据，即比例型数据是由大于等于0的实数组成的数据集合，而间隔型数据中可能出现负值。如某井各地层厚度数据，其差值表示两个地层的厚度差，比值反映了其中一个地层厚度是另一个地层厚度的百分之多少。

比例型数据所反映的数据概念最完整、意义最明确，因而是最重要的一类数据。

(二) 综合数据

综合数据是指由定量数据(或经定量化处理后的定性数据)，经有限次算术运算后得到的具有明确地质意义的综合性数据，例如总烃含量、时间—温度指数 *TTI* 等。另外，随机变量的各种数值特征，如平均值、标准差、极差、相关系数等都可认为是综合数据。

(三) 经验数据

经验数据是指在大量研究了地质现象和规律后，经过归纳或根据经验公式计算而得到的经验值。它们通常是大量地质信息的综合反映。经验数据的地质意义往往是十分明确的，但经验数据受到哪些地质因素的影响，以什么方式影响，经验数据和地质因素之间的数学关系是什么，这些问题往往是不确定的或不清楚的。

石油资源评价工作中经常使用经验数据，如单储系数、聚集系数、排烃系数等等。由于每个地质研究人员工作经历的局限性，经验数据往往具有较明显的地域性特征。因此，使用经验数据时要特别注意对比地质条件的相似性。不加选择地引用将导致错误的结果。

三、地质数据的特点

由于地质系统、地质条件和地质作用的复杂多变，各种技术测试手段之间的较大差异等原因，造成了地质数据本身的许多特点，主要包括以下几个方面：

① 地质数据的类型多，性质不一，反映的地质内容十分广泛，数量的多少和数据的精度相差悬殊，量纲变化大。

② 地质数据往往反映了多种地质因素综合作用的结果，具有混合分布特征。

③ 定量数据仍是地质数据的主要类型，对地质定性数据的定量化研究和应用尚不成熟。

上述特点说明了地质数据不是单一性质的集合，而是属于具有多种来源的复杂数据集合，这些特点是客观存在和不易改变的。使用地质数据时要特别注意其适用性，对不同的使用目的要选用不同的数据，同时还要加强和改进数据的加工和处理技术，只有这样才能有效使用地质数据，使数学地质方法取得较好的地质效果。

四、地质数据的误差

任何的观测手段都不可能得到与实际情况完全吻合的观测值。这是因为在野外观测、样品采集、管理、分析化验、仪器读数、资料整理过程中，由于受到工作人员的主观因素、仪器的精度限制、周围环境或随机因素、人为过失等的影响，必然使观测数据产生误差。误差是衡量地质数据质量的重要标志，按性质可分为三种类型。

1. 随机误差

随机误差是指在观测或测量过程中由不可控制的、无规律的偶然因素引起的误差，一般近似服从均值为0的正态分布。这类误差的大小和正负各不相同，当观测次数增加时，误差均值

趋于0。随机误差导致观测数据在一定范围内出现波动，称为观测数据的波动性或统计涨落性。

2. 系统误差

系统误差是指由观测系统本身所引起的误差。如仪器不准确，测量方法不合理，测量条件或环境的不同，观测者的不同习惯等因素引起的误差都是系统误差。这类误差往往使观测数据整体上偏大或偏小，可以用一定的手段校正观测数据，降低这类误差。

3. 过失误差

过失误差是指在数据观测和数据整理过程中，受到各种干扰和人为过失等因素影响所产生的误差。这类误差使地质数据失去了真实性和代表性，称为失真。如样品的污染，仪器的瞬时故障，数据整理过程中的笔误等都可能使观测数据失真。过失误差的多少和大小一般反映了观测人员的水平高低，所形成的失真数据往往是难以校正的，它对数据处理结果会产生极其不利的影响。

五、数据矩阵

地质数据的数量一般比较多，出于数据处理的需要，可将地质数据用数据矩阵表示。矩阵的每一列是一个地质变量的多个观测值，每一行是包含多个地质变量观测值的一个样品。如果地质数据包含 m 个变量的 n 次观测值（n 个样品），则可用下列 n 行 m 列的数据矩阵 X 表示：

$$X = [x_{ij}]_{n\times m} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

上述数据矩阵中包括了对 m 个地质变量的 n 次观测值，x_{ij}表示第 j 个地质变量的第 i 次观测值。由于数学习惯的不同，也可能用每列表示一个样品，每行表示一个地质变量的多个观测值。这样，上述数据矩阵成为下列形式：

$$X = [x_{ij}]_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

其中 x_{ij}表示第 i 个地质变量的第 j 次观测值。对于不同的数据矩阵表示方式，在处理时应注意区别对待。本书中一般按第一种方式表示数据矩阵。

例如，某探区已发现5个地质圈闭，为了描述这些圈闭的地质特征，选用了圈闭面积、闭合高度、长短轴比、埋藏深度共4项地质变量，5次观测数据如表2-1所示。

表2-1　地质圈闭数据表

地质变量 / 圈闭编号	圈闭面积/10^2m^2	闭合高度/m	长短轴比	埋藏深度/m
1	1000	500	1.5	2000
2	250	150	1.0	2200
3	100	70	3.0	1500
4	10	200	2.0	1800
5	40	100	5.0	2500

将上述数据整理为下列5行4列的数据矩阵：

$$X = [x_{ij}]_{5\times4} = \begin{bmatrix} 1000 & 500 & 1.5 & 2000 \\ 250 & 150 & 1.0 & 2200 \\ 100 & 70 & 3.0 & 1500 \\ 10 & 200 & 2.0 & 1800 \\ 40 & 100 & 5.0 & 2500 \end{bmatrix} \tag{2-1}$$

§2 地质数据的预处理

由于观测数据的量纲不同及存在各种误差等原因，将原始地质观测数据直接用于计算往往是不合适的。因此在进行正式计算之前需要对观测数据进行预处理，形成供进一步计算使用的数据。地质数据的预处理是定量计算过程中不可缺少的一个重要环节，它已成为数学地质的重要内容之一。

一、定量数据的标准化

不同地质变量原始观测值的单位、量纲以及数值大小、变化范围是不相同的，如果对原始数据直接使用，可能突出观测值较大地质变量的作用，降低观测值较小地质变量的作用。为克服数据中存在的这种不合理现象，在进行计算之前要将各地质变量的观测值变换到某种规范尺度之下，即定量数据的标准化。严格讲，定量数据的标准化包括对变量和样品观测值的标准化，但一般情况下只考虑对变量观测数据的标准化，少数情况下使用对样品的标准化。另外，为计算方便，定量数据的标准化一般在数据矩阵的基础上进行。

1. 总和标准化

总和标准化是指将变量的每个观测值变换为它与该项变量所有观测值总和的比值。因此，在变换后，数据矩阵的元素值在[0,1]之间，且每个变量的所有观测值之和等于1。具体变换公式为：

$$x'_{ij} = \frac{x_{ij}}{x_{\cdot j}} \quad (i = 1,2,\cdots,n \quad j = 1,2,\cdots,m) \tag{2-2}$$

式中 x'_{ij}—— 变换后的数据；

x_{ij}—— 变换前的数据；

$x_{\cdot j}$—— 第 j 个变量观测值总和，$x_{\cdot j} = \sum_{k=1}^{n} x_{kj}$，$n$ 为样品总数。

对式(2-1)中所列数据矩阵按式(2-2)进行变换，计算出各列(变量观测值)的总和如下：

$x_{\cdot 1}=1400, x_{\cdot 2}=1020, x_{\cdot 3}=12.5, x_{\cdot 4}=10000$

变换后的数据矩阵如下：

$$X' = \begin{bmatrix} 0.714 & 0.049 & 0.120 & 0.200 \\ 0.179 & 0.147 & 0.080 & 0.220 \\ 0.071 & 0.069 & 0.240 & 0.150 \\ 0.007 & 0.196 & 0.160 & 0.180 \\ 0.029 & 0.098 & 0.400 & 0.250 \end{bmatrix}$$

当样品数只有二个时，我们可以把每个变量变换后的观测值看成是二维平面上的点，第一个样品中各变量观测值作为 x 坐标，第二个样品中的变量观测值作为 y 坐标。由于进行总和

标准化后的变量观测值之和为1，因此变换后的点必然落在直线 $x+y=1$ 上。若将变换前的原始点和坐标原点连成一条直线，可以证明，该直线和 $x+y=1$ 的交点就是总和标准化后的点。

例如，有2个样品，5个变量组成的数据矩阵如下：

$$X=\begin{bmatrix}0.25 & 0.25 & 0.70 & 0.75 & 0.35\\1.00 & 0.50 & 0.80 & 0.50 & 0.10\end{bmatrix}$$

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}0.20 & 0.33 & 0.47 & 0.60 & 0.78\\0.80 & 0.67 & 0.53 & 0.40 & 0.22\end{bmatrix}$$

变换前后观测值点的空间分布如图2-1所示。

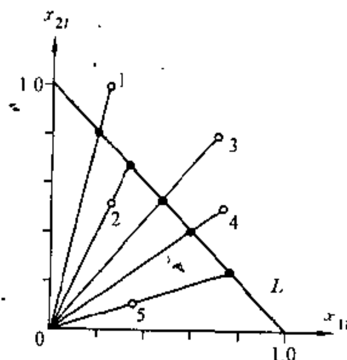


图2-1 二个样品的总和标准化

2. 最大值标准化

最大值标准化是将每个变量的观测值除以该变量所有观测值中的最大者。进行最大值标准化后的观测值中最大值为1。具体变换公式为：

$$x'_{ij}=\frac{x_{kj}}{\max\limits_{1\leqslant k\leqslant n} x_{ij}} \qquad (i=1,2,\cdots,n \quad j=1,2,\cdots,m) \tag{2-3}$$

式中 x'_{ij}——变换后的数据；

x_{ij}——变换前的数据；

$\max\limits_{1\leqslant i\leqslant n} x_{ij}$——第 j 个变量观测值中的最大者。

对式(2-1)中所列数据矩阵按式(2-3)进行变换，1至4列(变量观测值)的最大值如下：

1000，500，5，2500

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}1.00 & 1.00 & 0.30 & 0.80\\0.25 & 0.30 & 0.20 & 0.88\\0.10 & 0.14 & 0.60 & 0.60\\0.01 & 0.40 & 0.40 & 0.72\\0.04 & 0.20 & 1.00 & 1.00\end{bmatrix}$$

当样品数只有二个时，可将每个变量变换后的观测值看成是二维平面上的点，由于最大值标准化后每个点必有一个坐标值为1，所以变换后点必然落在直线 $x=1$ 或 $y=1$ 上。

例如，有2个样品，5个变量组成的数据矩阵如下：

$$X=\begin{bmatrix}0.25 & 0.25 & 0.70 & 0.75 & 0.35\\1.00 & 0.50 & 0.80 & 0.50 & 0.10\end{bmatrix}$$

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}0.25 & 0.50 & 0.86 & 1.00 & 1.00\\1.00 & 1.00 & 1.00 & 0.67 & 0.26\end{bmatrix}$$

变换前后观测值点的空间分布如图2-2所示。

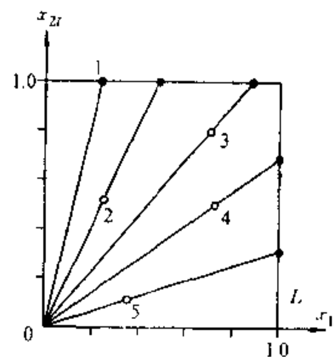


图2-2 二个样品的最大值标准化

3. 模标准化

若将每个变量的观测值视为 n 维空间上的向量，则模标准化是将每个变量的观测值除以

$$x'_{ij}=\frac{x_{ij}}{|\vec{X}_j|}\quad (i=1,2,\cdots,n\quad j=1,2,\cdots,m) \tag{2-4}$$

式中 x'_{ij}——变换后的数据；

x_{ij}——变换前的数据；

$|\vec{X}_j|$——第 j 个变量观测值向量的模，$|\vec{X}_j|=\sqrt{\sum_{k=1}^{n}x_{kj}^2}\quad (j=1,2,\cdots,m)$。

对式(2-1)中所列数据矩阵按式(2-4)进行变换，1 至 4 列(变量观测值)的模如下：

$$|\vec{X}_1|=1036.436,|\vec{X}_2|=572.189,|\vec{X}_3|=6.423,|\vec{X}_4|=4536.518$$

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}0.965 & 0.847 & 0.234 & 0.441\\ 0.241 & 0.262 & 0.156 & 0.485\\ 0.096 & 0.122 & 0.467 & 0.331\\ 0.010 & 0.350 & 0.311 & 0.397\\ 0.039 & 0.175 & 0.778 & 0.551\end{bmatrix}$$

当样品数只有二个时，将每个变量变换后的观测值看成是二维平面上的点，由于模标准化后每个点 x、y 坐标的平方和为 1，所以必然落在圆 $x^2+y^2=1$ 上。

例如，有 2 个样品，5 个变量组成的数据矩阵如下：

$$X=\begin{bmatrix}0.25 & 0.25 & 0.70 & 0.75 & 0.35\\ 1.00 & 0.50 & 0.80 & 0.50 & 0.10\end{bmatrix}$$

变换后的数据矩阵如下：

$$X=\begin{bmatrix}0.243 & 0.447 & 0.659 & 0.832 & 0.962\\ 0.970 & 0.894 & 0.753 & 0.555 & 0.275\end{bmatrix}$$

变换前后观测值点的空间分布如图 2-3 所示。

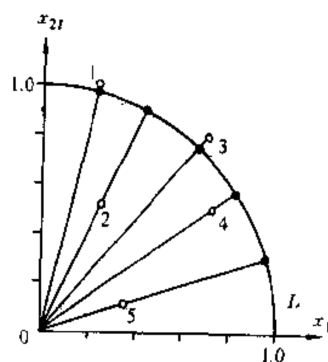


图 2-3 二个样品的模标准化

4. 中心标准化

中心标准化是将变量的每个观测值减去该变量所有观测值的平均值。变换后所有观测值的平均值等于 0，即总和为 0。具体变换公式为：

$$x'_{ij}=x_{ij}-\bar{x}_j\qquad (i=1,2,\cdots,n\quad j=1,2,\cdots,m) \tag{2-5}$$

式中 x'_{ij}—— 变换后的数据；

x_{ij}—— 变换前的数据；

$\bar{x}_j$—— 第 j 个变量观测值的平均值，$\bar{x}_j=\frac{1}{n}\sum_{k=1}^{n}x_{kj},j=1,2,\cdots,m$。

对式(2-1)中所列数据矩阵按式(2-5)进行变换，1 至 4 列(变量观测值)的平均值如下：

$$\bar{x}_1=280,\bar{x}_2=204,\bar{x}_3=2.5,\bar{x}_4=2000$$

变换后的数据矩阵如下：

$$X' = \begin{bmatrix} 720 & 296 & -1.0 & 0 \\ -30 & -54 & -1.5 & 200 \\ -180 & -134 & 0.5 & -500 \\ -270 & -4 & -0.5 & -200 \\ -240 & -104 & 2.5 & 500 \end{bmatrix}$$

当样品数只有二个时，将每个变量变换后的观测值看成是二维平面上的点，由于中心标准化后每个点 x、y 坐标之和为 0，所以变换后点必然落在直线 $x+y=0$ 上。

例如，有 2 个样品，5 个变量组成的数据矩阵如下：

$$X = \begin{bmatrix} 0.25 & 0.25 & 0.70 & 0.75 & 0.35 \\ 1.00 & 0.50 & 0.80 & 0.50 & 0.10 \end{bmatrix}$$

变换后的数据矩阵如下：

$$X' = \begin{bmatrix} -0.375 & -0.125 & -0.050 & 0.125 & 0.125 \\ 0.375 & 0.125 & 0.050 & -0.125 & -0.125 \end{bmatrix}$$

变换前后观测值点的空间分布如图 2-4 所示。

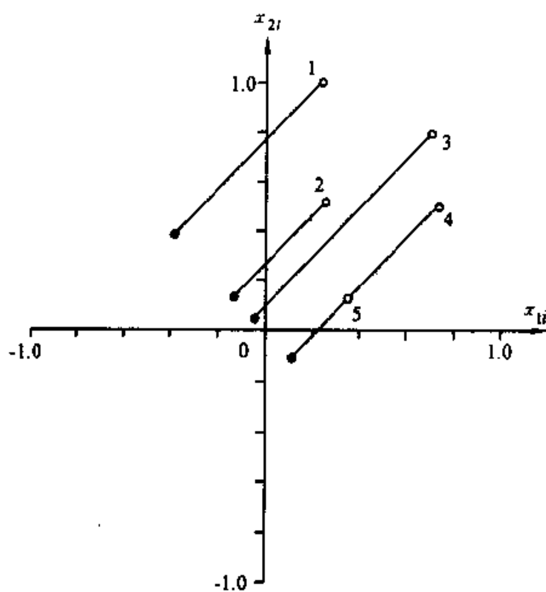


图 2-4 三个样品的中心标准化

5. 标准差标准化

标准差标准化是将变量的每个观测值减去该变量所有观测值的平均值，再除以该变量观测值的标准差。每个观测值变量变换后的平均值等于 0，标准差均为 1。这是一种常用的标准化方法，变换后的数据又称为规格化数据。具体变换公式为：

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

$$(i = 1,2,\cdots,n\ j = 1,2,\cdots,m) \quad (2\text{-}6)$$

式中 x'_{ij}—— 变换后的数据；

x_{ij}—— 变换前的数据；

$\bar{x}_j$—— 第 j 个变量观测值的平均值，$\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}, j = 1,2,\cdots,m$；

s_j—— 第 j 个变量的标准差，$s_j = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_{kj} - \bar{x}_j)^2}, j = 1,2,\cdots,m$。

对式(2-1)中所列数据矩阵按式(2-6)进行变换，1 至 4 列(变量观测值)的平均值如下：

$$\bar{x}_1 = 280, \bar{x}_2 = 204, \bar{x}_3 = 2.5, \bar{x}_4 = 2000$$

1 至 4 列(变量观测值)的标准差如下：

$$s_1 = 369.378, s_2 = 154.480, s_3 = 1.414, s_4 = 340.588$$

变换后的数据矩阵如下：

$$X' = \begin{bmatrix} 1.949 & 1.916 & -0.707 & 0.000 \\ -0.081 & -0.350 & -1.016 & 0.587 \\ -0.487 & -0.867 & 0.354 & -1.468 \\ -0.731 & -0.026 & -0.354 & -0.587 \\ -0.650 & -0.673 & 1.768 & 1.468 \end{bmatrix}$$

当样品数只有二个时，将每个变量变换后的观测值看成是二维平面上的点，由于标准差标准化后每个点 x、y 坐标之和为 0，所以变换后点必然落在直线 $x+y=0$ 上，而且落在以原点为圆心，以 $\sqrt{2}$ 为半径的圆 $x^2+y^2=2$ 上。

例如，有 2 个样品，5 个变量组成的数据矩阵如下：

$$X = \begin{bmatrix} 0.25 & 0.25 & 0.70 & 0.75 & 0.35 \\ 1.00 & 0.50 & 0.80 & 0.50 & 0.10 \end{bmatrix}$$

变换后的数据矩阵如下：

$$X' = \begin{bmatrix} -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 \end{bmatrix}$$

变换前后观测值点的空间分布如图 2-5 所示。

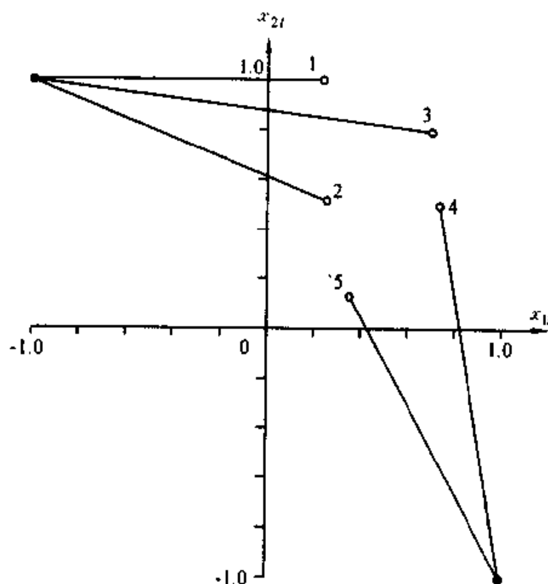


图 2-5　二个样品的标准差标准化

6. 极差标准化

极差标准化是将变量的每个观测值减去该变量所有观测值的平均值，再除以变量观测值的极差。对每个变量观测值变换后，极差均为 1。具体变换公式为：

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\max\limits_{1\leqslant k\leqslant n} x_{kj} - \min\limits_{1\leqslant k\leqslant n} x_{kj}}$$

$(i = 1,2,\cdots,n\ j = 1,2,\cdots,m)$ (2-7)

式中　x'_{ij}—— 变换后的数据；

x_{ij}—— 变换前的数据；

$\bar{x}_j$—— 第 j 个变量观测值的平均值，$\bar{x}_j = \frac{1}{n}\sum\limits_{k=1}^{n} x_{kj}$，$j = 1,2,\cdots,m$。

对式(2-1)中所列数据矩阵按式(2-7)进行变换，1 至 4 列(变量观测值)的平均值如下：

$$\bar{x}_1=280,\bar{x}_2=204,\bar{x}_3=2.5,\bar{x}_4=2000$$

1 至 4 列(变量观测值)的极差分别为：990，430，4，1000

变换后的数据矩阵如下：

$$X' = \begin{bmatrix} 0.727 & 0.688 & -0.250 & 0.000 \\ -0.030 & -0.126 & -0.375 & 0.200 \\ 0.182 & -0.312 & 0.125 & -1.500 \\ -0.273 & -0.009 & -0.125 & -0.200 \\ -0.242 & -0.342 & 0.625 & 0.500 \end{bmatrix}$$

当只有二个样品时，将每个变量变换后的观测值看成是二维平面上的点，由于极差标准化后每个点 x、y 坐标之差的绝对值为 1，所以变换后的点必然落在直线 $x-y=1$ 或 $x-y=-1$ 上。另外从式(2-7)可推知，x、y 的值取 0.5 和 -0.5 其中之一。例如，有 2 个样品，5 个变量组成的数据矩阵如下：

$$X=\begin{bmatrix}0.25 & 0.25 & 0.70 & 0.75 & 0.35\\1.00 & 0.50 & 0.80 & 0.50 & 0.10\end{bmatrix}$$

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}-0.5 & -0.5 & -0.5 & 0.5 & 0.5\\0.5 & 0.5 & 0.5 & -0.5 & -0.5\end{bmatrix}$$

变换前后观测值点的空间分布如图 2-6 所示。

7. 极差正规化

极差正规化是将变量的每个观测值减去该变量所有观测值的最小值，再除以变量观测值的极差。变换后每个变量观测值在 0～1 之间，最大为 1，最小为 0。这是一种常用的标准化方法。具体变换公式为：

$$x'_{ij}=\frac{x_{ij}-\min\limits_{1\leqslant k\leqslant n} x_{kj}}{\max\limits_{1\leqslant k\leqslant n} x_{kj}-\min\limits_{1\leqslant k\leqslant n} x_{kj}}$$

$(i=1,2,\cdots,n\ j=1,2,\cdots,m)$(2-8)

式中 x'_{ij}——变换后的数据；

x_{ij}——变换前的数据。

对式(2-1)中所列数据矩阵按式(2-8)进行变换，1 至 4 列(变量观测值)的最小值分别为：

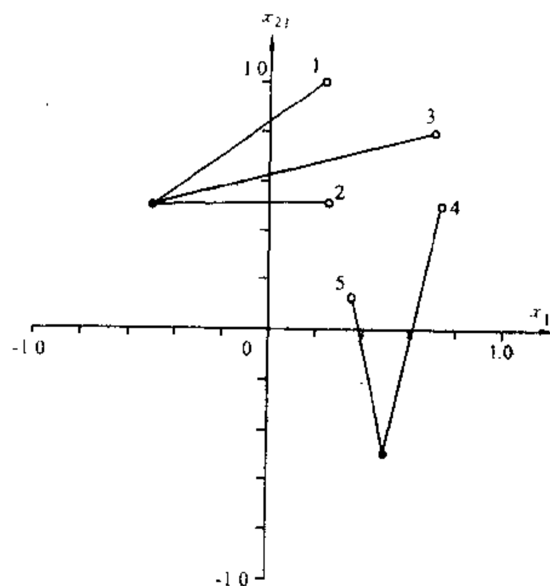


图 2-6 二个样品的极差标准化

10,70,1,1500

1 至 4 列(变量观测值)的最大值分别为：

1000,500,5,2500

变换后的数据矩阵如下：

$$X'=\begin{bmatrix}1.000 & 1.000 & 0.125 & 0.500\\0.242 & 0.186 & 0.000 & 0.700\\0.091 & 0.000 & 0.500 & 0.000\\0.000 & 0.302 & 0.250 & 0.300\\0.030 & 0.070 & 1.000 & 1.000\end{bmatrix}$$

当样品数只有二个时，将每个变量变换后的观测值看成是二维平面上的点，由于极差准化后每个点 x、y 坐标在 0～1 之间，而且其中必有 0 和 1，所以变换后点必然是(1，0)或(0，1)。

例如，有 2 个样品，5 个变量组成的数据矩阵如下：

$$X=\begin{bmatrix}0.25 & 0.25 & 0.70 & 0.75 & 0.35\\1.00 & 0.50 & 0.80 & 0.50 & 0.10\end{bmatrix}$$

变换后的数据矩阵如下：

$$X' = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

变换前后观测值点的空间分布如图 2-7 所示。

极差正规化是将变量观测值变换到闭区间[0，1]内。对一些特殊情况(如计算机绘图)，往往需要将数据变换到任意闭区间[a，b]，类似于极差正规化，可采用以下变换公式：

$$x'_{ij} = a + (b-a)\frac{x_{ij} - \min\limits_{1\leqslant k\leqslant n} x_{kj}}{\max\limits_{1\leqslant k\leqslant n} x_{kj} - \min\limits_{1\leqslant k\leqslant n} x_{kj}} \quad (i=1,2,\cdots,n \quad j=1,2,\cdots,m) \tag{2-9}$$

式中 x'_{ij}——变换后的数据；

x_{ij}——变换前的数据。

二、定性数据的定量化变换

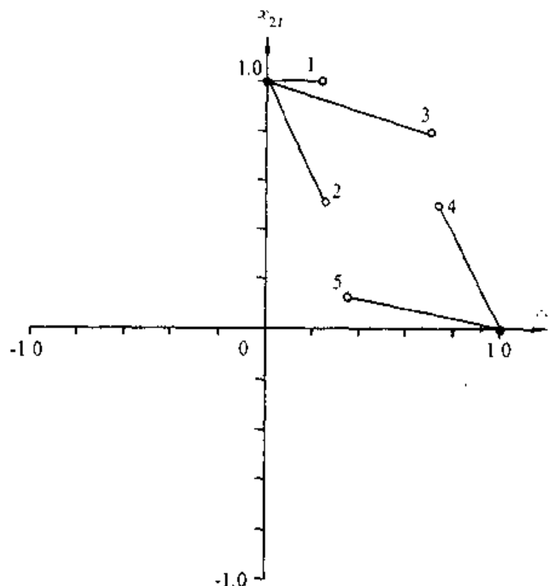


图 2-7 二个样品的极差正规化

地质数据中有一些数据属于定性数据，在地质研究工作中也要经常用到定性数据，有时甚至是不可避免的。例如，生油岩的颜色，钻井取出的岩心或岩屑的含油状态，地层中某种古生物化石的有无等等都是不能用数值描述的定性数据。定性数据不能直接参加运算，为了使定性数据能够用于定量研究，必须将定性数据的符号或代码赋以定量的数值，这就是对定性数据的定量化变换。近年来，对定性数据的定量处理，虽然已引起数学地质界的重视，但研究的深度还很不够。

1. 二态定性数据的变换

如果名义型数据或有序型数据只有两种对立状态，则称为二态定性数据。两种对立状态是指在两种仅有的状态中必须是其中的一种，亦即所谓"非此即彼"。在这种情况下，对定性数据采用 0,1 变换是十分方便的。

二态定性数据的 0,1 变换，其具体作法是把两种对立状态中的一种状态赋值为 1，另一种状态赋值为 0。通常是将对所研究问题有利的或肯定的状态赋值为 1，将不利的或否定的状态赋值为 0。例如，进行某一地区的地层对比时，同一层位的不同观测点，有时可找到对分层有意义的某种古生物化石，有时却找不到这种化石，在进行数据变换时，对出现化石的地点可赋值为 1、不出现的地点可赋值为 0。又如，在钻井过程中进行岩屑录井时，对于出现油砂的层位可赋值为 1，不出现油砂的层位可赋值为 0。在一般的情况下：

状态	有利或肯定状态	不利或否定状态
赋值	1	0

需要指出，0,1 变换是一种简单而实用的变换方法，而且变换后的数据 0,1 可与极差正规化变换后的定量数据混合使用。前已述及，经极差正规化变换后的定量数据，其特征是所有数

据均被压缩到[0,1]闭区间范围内，因此，经0,1变换后的定性数据可看作是定量数据的两种极端状态。可见，这种变换方法是符合实际情况的。

2. 有序型多态定性数据的变换

如果定性数据的状态数大于2，而且状态可按一定的次序进行排列，这种数据可称为有序型多态定性数据。例如，从钻井中所取出的岩心，按其含油程度可分为如下四个级别，即：

状态	不含油	油斑	含油	饱含油
赋值	0	1	2	3

对有序型多态定性数据进行变换时，一般是用非负整数进行赋值，由最低级的状态到最高级的状态，其赋值要逐渐增大。根据实际情况，可采用等差式的等级赋值，也可以采用非等差式的等级赋值。上述对岩心含油程度的赋值就是等差式的等级赋值，等差为1。另外如，根据泥岩的颜色划分生油条件时，采用下面的赋值方法：

状态	红色	浅灰色	灰色	黑色
赋值	0	1	3	6

这种赋值就是非等差式的赋值方法。

有序型多态定性数据经过赋值后，如果再按前一节中定量数据的变换方法作进一步的变换，则变换后的多态定性数据可与定量数据混合使用。

三、原始数据的均匀化、简缩和增补

1. 原始数据的均匀化(网格化)

在计算机绘图等工作中，往往需要对数据进行网格化处理，即将地质数据分配到一些规则的矩形网格交点(网格点)上，进行网格化的地质数据一般是在平面上分布的、和位置有关的定量数据(比例型数据)，如一个地区许多井中某地层厚度数据，每口井都有一个地理坐标，每个数据之间都具有一个实际的平而距离。网格化的方法很多，这里仅介绍一种简单而又常用的网格化方法：象限距离加权平均法。

在以某一个网格点为坐标原点的坐标系的四个象限中，各选一个距该点最近的数据点，假设其平面距离值分别为：r_1, r_2, r_3, r_4，相应的数据值分别为：z_1, z_2, z_3, z_4，考虑到距离越小，对该网格点的影响越大，因此取距离 $r_i(i=1,2,3,4)$的倒数作为权。

该网格点的数据 z 可按下列公式进行预测：

$$z=\sum_{i=1}^{4}\frac{z_i}{r_i}\Bigg/\sum_{i=1}^{4}\frac{1}{r_i} \tag{2-10}$$

对于某些网格点(如边界网格点)，不能在四个象限中都找到数据点，则在有数据点的象限中取距离最近点进行加权平均，这时被加权平均的数据个数小于4个，但至少是1个。另外，也可在每个象限中选取多个离散数据点进行加权平均。过程与上类同。

对每个网格点进行上述计算，即可完成对数据的网格化工作。处理过程中要注意 $r_i=0$ 的情况，此时网格点上的数据与 z_i 相同。进行数据网格化的方法很多，如全点插值、圆内插值、曲面插值、克里金法等，在此不加赘述。

2. 原始数据的简缩

当分布于区域上的地质数据样品数量很多时，或者是数据在区域上的分布极不均匀时，可能出现反映相同地质特征的多个近似样品，这不仅会使计算量大大增加，而且无助于最终的成

果解释，甚至在计算过程中还会出现不可预料的计算病态问题。因此，需要对那些作用不大或相近、可有可无的多余数据予以舍弃，这就是数据的简缩。

数据的简缩方法一般包括分区加权平均法、分区滑动平均法和随机删点法。

(1) 分区加权法

假如在一个探区中有 N 个地质数据，则可根据实际需要将探区分成大小相等或不相等的 m 个小区，要求在每个小区中至少有一个原始数据点。如果其中第 j 个小区中有 $n_j(j=1,2,\cdots,m)$ 个数据点，那么有：

$$N = n_1 + n_2 + \cdots + n_j + \cdots + n_m$$

令第 j 个小区中每个数据点的权为 $1/n_j$，则每个小区中所包含的数据点的权和等于 1。而且有：

$$n_1 \frac{1}{n_1} + n_2 \frac{1}{n_2} + \cdots + n_j \frac{1}{n_j} + \cdots + n_m \frac{1}{n_m} = m$$

这样一来，按加权平均法在每个小区都可计算出一个综合数据点(重心)，从而将原来数据量很大的地质数据简化为 m 个有效数据点。在随后的计算中只要用 m 个有效数据点就可以了。

地质数据经常是多变量观测数据，如果每个数据由 p 个地质变量观测值组成，则第 j 个小区第 k 个地质变量观测值的简缩值可用式(2-11)计算，即：

$$z_{jk} = \frac{1}{n_j}\sum_{i=1}^{n_j} z_{jki} \quad (j = 1,2,\cdots,m \quad k = 1,2,\cdots,p) \tag{2-11}$$

式中 z_{jk}——第 j 个小区第 k 个地质变量的简缩观测值；

n_j——第 j 个小区中地质数据个数；

z_{jki}——第 j 个小区中第 k 个地质变量观测值的第 i 个数据。

(2) 分区滑动平均法

分区滑动平均法与分区加权平均法一样，也要将研究区分成若干个小区，分区原则二者相同，但是，分区滑动平均法要考虑简缩后数据点的位置。

如果第 j 个小区中有 $n_j(j=1,2,\cdots,m)$ 个数据点，每个数据点含 p 个地质变量的观测值，其中第 i 个数据点的坐标为 (x_{jki}, y_{jki})，变量观测值为 z_{jki}。第 j 个小区简缩后的有效数据点的坐标值及变量值可用下面的(2-12)、(2-13)、(2-14)式计算求出：

$$x_{jk} = \sum_{i=1}^{n_j} x_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \tag{2-12}$$

$$y_{jk} = \sum_{i=1}^{n_j} y_{jki} \cdot z_{jki} \Big/ \sum_{i=1}^{n_j} z_{jki} \tag{2-13}$$

$$z_{jk} = \sum_{i=1}^{n_j} z_{jki} / n_j \tag{2-14}$$

$(j = 1,2,\cdots,m \quad k = 1,2,\cdots,p)$

式中 x_{jk}、y_{jk}——第 j 个小区第 k 个地质变量观测值简缩后的横坐标与纵坐标；

z_{jk}——第 j 个小区第 k 个地质变量的简缩值；

x_{jki}、y_{jki}——第 j 个小区第 k 个地质变量观测值的第 i 个数据点的横坐标与纵坐标；

z_{jki}——第 i 个小区第 k 个地质变量观测值的第 i 个数据；

n_j——第 j 个小区中的地质数据个数。

按上述公式计算的坐标可能有 p 个，若需要一个统一的坐标点，则可根据地质变量的作用大小，采用加权平均法计法得出。另外，根据实际情况，也可采用其它的计算方法。

（3）随机删点法

如果探区中某些局部地区的数据点过密，则可以随机删去一些数据点，这不仅可以减少计算工作量，也可以保证计算过程的稳定性。随机删点时，可以人为地随机删除一些数据点，也可以对数据点顺序编号，按随机数抽样法进行删除，抽样方法见本书第十章。

3. 数据的增补

一般情况下，探区内投入的勘探工作量是不均匀的，特别是早期勘探阶段。因此所获得的数据在区域上会出现空白区，这种情况下往往需要补充数据，这就是数据的增补。

在没有数据点的大片空白地区，为了补充一些数据点，可以用临近已有的地质数据进行外推，即根据数据的变化趋势，补充一些数量适当的数据点，或者根据临近地区的已有数据点，用某种约定的插值方法补充一些数据点。必须注意的是，补点的目的是为了全区计算上的稳定，补点后原空白区的计算结果是不可信的。

此外，由于探区中各处取样点的化验分析项目不完全一致，因而使多变量的数据矩阵中，某些样品缺少一些变量的观测值。但是，由于研究工作上的需要又不能删掉这些缺少数值的变量。为了增补缺项位置上的变量观测值，一般情况下可用该变量已有观测数据的平均值代替或者用区域上临近数据的平均值代替。

四、离群数据的判定及处理

由于各种原因造成的观测数据离群（局部异常高值和异常低值）现象，往往直接影响到基于观测数据的计算过程以及对计算结果的合理解释。对某些已知因素造成的数据离群，可进行相应的数据校正。如在油气地表化探中，由地表自然地理条件、土壤的类型、颜色等所导致的有关指标含量的差异，可进行相应的校正，消除数据中的已知干扰因素。

如果离群数据是对地质情况的真实反映，但这些离群数据对计算过程及结果会产生一些消极影响的话，对离群数据进行适当的处理应该是必要的。如果离群数据根本就是人为等因素所导致的错误数据，则应该毫不犹豫进行删除或重新进行观测。然而，判断离群数据出现的原因是很困难的，实际工作中总是假设数据是真实的，我们仅在数据真实的前提下讨论对离群数据的挑选和处理问题。

对离群数据进行处理的第一步是挑选离群数据，这里涉及到离群数据的界限问题。下面就离群数据的界限确定和处理方法进行简单介绍。

1. 类比法

以实际工作经验确定一个离群数据的界限，以此判定是否存在离群数据。B. N. 斯米尔诺夫根据实际经验，总结出一个确定矿床品位离群数据的界限，见表 2-2。

表 2-2　矿床品位离群数据的界限

矿床类型	组分分布性质	典　型　矿　床	离群品位高出平均品位的倍数
Ⅰ	很均匀	大多数沉积矿床	2～3
Ⅱ	均匀	复杂沉积矿床与变质矿床	4～5
Ⅲ	不均匀	绝大多数有色金属矿床	8～10
Ⅳ	很不均匀	大多数稀有金属矿床和金矿床	12～15
Ⅴ	极不均匀	某些稀有金属矿床和金矿床	>15

表 2-2 的离群品位高出平均品位的倍数是一些经验数据，只能作为参考。从矿床成因角度

看，绝大多数的油气藏都属于与沉积岩有关的矿床。所以，确定油气勘探、开发中有关地质离群数据界限时，可以参照表 2-2 中的 Ⅰ、Ⅱ 矿床类型。

2. 计算法

利用一个经验公式确定离群数据的界限。H. B. 沃洛多莫夫给出了下面的公式，通过计算来确定离群数据的界限。

$$ch = c_1 + (n-1)c_1 M = c_1 + \frac{(n-1)c_1(c_1-c_2)}{c_2} \tag{2-15}$$

式中 ch——正常数据的最高值，即大于 ch 的数据则为离群数据；

c_1——校正前（包括离群数据）的样品平均值；

c_2——校正后（不包括离群数据）的样品平均值；

n——包括离群数据在内的样品总数；

$M=(c_1-c_2)/c_2$

离群数据在一组数据中，一般只有少数几个，个数太多时则已不是离群数据。在实际计算时，令 $M=20\%\sim30\%$，由式(2-15)可计算出离群数据的界限值。

这种方法计算出的 ch 值显然与子样容量 n 有关，n 越大，ch 值的偏离可能越大。所以只适合于对小子样的检验。

3. 统计检验法

在观测数据来自一个正态总体的前提下，方法的思路是检验数据是否服从正态分布。若检验通过，则认为数据中不存在离群数据，否则认为数据中存在离群数据，这时应判定出其中的离群数据并进行有关处理。进行检验的关键是寻求一个合适的数据统计量及所服从的分布，并在此基础上确定相应的假设检验方法。

(1) 正态分布 χ^2 检验法

对来自正态总体的 n 个观测数据：$X_i, i=1,2,\cdots,n$，将区间 $(-\infty,+\infty)$ 划分为 m 个小区间：

$$(a_0,a_1),(a_1,a_2),\cdots,(a_{m-1},a_m), \qquad a_0=-\infty, a_m=+\infty。$$

设 v_i 为数据落入第 i 个小区间 (a_{i-1},a_i) 的个数（频数），p_i 为数据落入第 i 个小区间 (a_{i-1},a_i) 的理论概率。

对于备择假设：H_0：观测数据来自正态总体

H_1：观测数据不来自正态总体

若 H_0 为真，由皮尔逊定理知：

$$\eta = \sum_{i=1}^{n} \frac{(v_i - np_i)^2}{np_i} \sim \chi^2_{m-3} \quad (m>3) \tag{2-16}$$

因此确定正态分布 χ^2 检验方法如下：

① 计算观测数据均值 $\overline{X}$ 和方差的极大似然估计量 S^2：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n}\sum_{1=1}^{n}(X_i - \overline{X})^2$$

② 数据在来自正态总体 $N(\overline{X},S^2)$ 的假设下，求落入小区间 (a_{i-1},a_i) 的频数 v_i 及理论概率 p_i，$i=1,2\cdots,m$，并计算统计量 η。

③ 给定检验水平 α，查表求取拒绝域临界值 $\chi^2_{m-3}(\alpha)$，若 $\eta<\chi^2_{m-3}(\alpha)$，则接受假设 H_0，即认为数据来自正态总体。否则拒绝 H_0，存在离群数据。

在对地质观测数据进行正态检验时，可将区间($\min\limits_{1\leqslant i\leqslant n}X_i$, $\max\limits_{1\leqslant i\leqslant n}X_i$)均匀划分为 m 个小区间：

$(a_0,a_1),(a_1,a_2),\cdots,(a_{m-1},a_m)$， $a_1=\min\limits_{1\leqslant i\leqslant n}X_i$， $a_{m-1}=\max\limits_{1\leqslant i\leqslant n}X_i$

令 $a_0=-\infty$, $a_m=+\infty$.则共形成 m 个区间。之后按上述①～③的步骤进行检验。区间的个数 m 随数据点的增加而增加，一般取 m 为10～40，检验水平 α 可取0.1、0.05、0.01等，相应临界值由 χ^2 分布临界值表中查取，如 $\chi^2_{30}(0.05)=43.773$ 等。

(2) 正态分布偏度、峰度检验法

随机变量 X 的偏度 E_p 、峰度 E_f 是指 X 的标准化变量 $\frac{X-\mu}{\sigma}$ 的三阶中心矩和四阶中心矩：

$$E_p=E\left(\frac{X-\mu}{\sigma}\right)^3 \qquad E_f=E\left(\frac{X-\mu}{\sigma}\right)^4$$

对于观测数据 $X_i,i=1,2,3,\cdots,n$ 。可计算出该批数据的偏度 E_p 和峰度 E_f 的矩估计：

$$U_p=\frac{U_3}{S^3} \qquad U_f=\frac{U_4}{S^4}$$

式中 $U_3=\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^3$ 样本三阶中心矩

$U_4=\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^4$ 样本四阶中心矩

$\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$ 样本数据均值

$S=\sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}$ 样本标准差

对于备择假设： H_0 ：观测数据来自正态总体

H_1 ：观测数据不来自正态总体

若 H_0 为真，则由概率统计理论可证，当 n 充分大时近似有：

$$U_p\sim N(0,\ S_p)\ ,\ U_f\sim N\left(3-\frac{6}{n+1},\ S_f\right)$$

$$S_p=\frac{6(n-2)}{(n+1)(n+3)}$$

$$S_f=\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

即

$$\frac{U_p}{\sqrt{S_p}}\sim N(0,1),\frac{U_f-3+\frac{6}{n+1}}{\sqrt{S_f}}\sim N(0,1)$$

因此确定正态分布偏度、峰度检验方法如下：

① 对观测数据 $X_i(i=1,2,\cdots,n)$ ，求出其偏度和峰度的矩估计 U_p , U_f 。

② 对给定的显著检验性水平 α ，求出检验拒绝域临界值：

$$P_p=Z_{\alpha/2}\cdot\sqrt{S_p} \qquad P_f=Z_{\alpha/2}\cdot\sqrt{S_f}+3-\frac{6}{n+1} \tag{2-17}$$

若 $|U_p|<P_p$ 且 $|U_f|<P_f$ ，则接受 H_0 ，认为该批数据服从正态分布，否则拒绝 H_0 ，认为该批数据不服从正态分布，存在离群数据。

检验的显著性水平 α 通常取值为0.1、0.05、0.01、0.005、0.001，一般以 $\alpha=0.05$ 、0.01为

最常用。当 $\alpha=0.01$ 时，$Z_{\alpha/2}=2.575$，当 $\alpha=0.05$ 时，$Z_{\alpha/2}=1.96$，其它可由标准正态分布表直接查取。

(3) 统计检验法中离群数据的界限

对随机变量 X，若其均值为 μ，方差为 σ^2，则由契比雪夫不等式知：

$$P(\mu-2\sigma<X<\mu+2\sigma)\geqslant 0.8889 \quad P(\mu \quad 3\sigma<X<\mu+3\sigma)\geqslant 0.9375 \tag{2-18}$$

若随机变量 $X\sim N(\mu,\sigma^2)$，则有：

$$P(\mu-2\sigma<X<\mu+2\sigma)=0.9544 \quad P(\mu-2\sigma<X<\mu+3\sigma)=0.9974 \tag{2-19}$$

因此，对于观测数据 $X_i(i=1,2,\cdots,n)$，可确定它以大概率落人的区间：

$$(X-2S,X+2S)\text{ 或 }(X-3S,X+3S)$$

若某个数据不在上述区间内，则认为该数据离群，需进行处理。否则认为该数据正常。

4. 对离群数据的处理

无论用哪种方法挑选出了离群数据，下一步均要对离群数据进行处理。一般常用三种方式：

① 将数据用相邻数据之插值代替，或用某种平均值代替。

② 将数据缩小或放大为上述区间的边界值，即对数据离群程度的抑制。

③ 将该数据剔除，即放弃该数据，数据总数减少。

采用哪种方法需结合数据的情况而定。

5. 方法实施步骤

以统计检验法为例，讨论对离群数据的处理步骤。由于某些地质观测数据离散程度较高，按前述方法对数据进行一次处理后，新的数据体可能仍不能满足正态分布的要求，这时，必须在新的数据体基础上再进行同样的处理，反复多次，直至数据体满足正态分布。因此，可将对离群数据的判定和处理归纳为一个迭代过程。

对来自正态总体的观测数据 $X_i(i=1,2,\cdots,n)$，处理步骤如下：

① 输人所有观测数据，给定检验水平 α。

② 在检验水平 α 下，进行该批数据的正态分布检验（χ^2 检验或偏度、峰度检验），若通过检验则结束，否则进行下一步。

③ 判定离群数据并进行相应处理，形成新的数据。

④ 重复②～③，直至数据体通过正态分布检验。

计算机程序设计流程见图 2-8。

§3 取样问题

由于地质问题所涉及的地域、空间十分广阔，因而在研究工作中，往往只能从总体中抽出一部分样品进行研究，即用一个子样来研究总体。但是，所抽取的子样能否代表总体？怎样取样才合理？这就需要研究取样问题。

一、随机取样

随机取样要求总体中每个样品被抽到的概率是相等的，每次取样是相互独立的。常用的方法两种。

1. 抽签法

抽签法是最直观的随机取样方法。例如，某个探区经过多年勘探已积累了大量的原油、天

然气、地下水的化验室分析数据，每种数据多达几千个，甚至上万个。如果想用随机取样法抽查某种化验项目的100个样品数据时，可将已有的分析数据逐一编上一个号码，并把每个号码写在卡片上，这些卡片经过充分混合后，随机取出100个，那么，卡片上编号对应的100个分析数据就是一个随机子样。

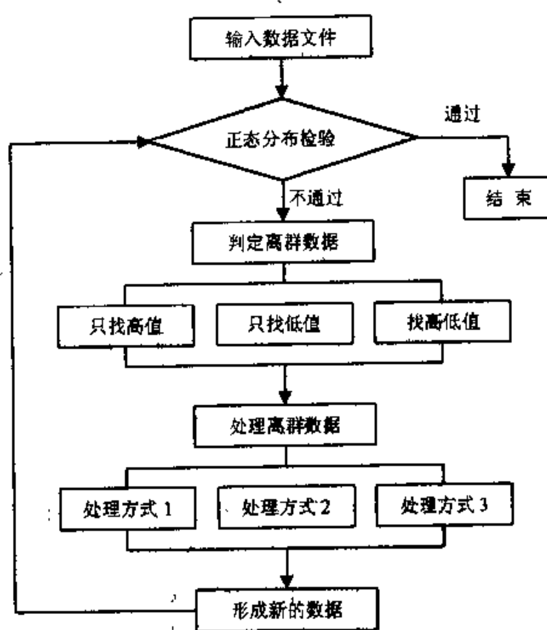


图 2-8 离群数据处理流程图

2. 随机数抽样法

许多数学手册上都附有随机数表，表中的数据之间无任何规律可循，从随机数表的任何一页、任何一行、任何一列的数字开始向上、向下、向左、向右读取位数相同的数字序列就是一组随机数。

如果某项地质数据的数量有 N 个，想从中随机抽取 M 个数据 $(M \ll N)$ 组成一个随机子样，可将 N 个数据逐一进行编号，从1开始编号到 N 为止。假如 N 是一个 P 位数，则可从随机数表中读取 M 个 P 位随机数。当随机数的值 R 大于等于1，并且小于等于 N 时，可以直接使用。当随机数的值 R 大于 N 时，需要取 R 除以 N 的余数。

例如，有1250个化验分析数据，即 $N=1250$，要从中随机抽取100个数据，组成一个随机子样，即 $M=100$。1250是4位数字，因而需要从随机数表中读100个4位随机数。下面列举的随机数是从某个随机数表第一页第5行第11列开始向右读，读完第5行时再从第6行第1列继续读，每个随机数是从所在的行列开始向下读4位，则得到如下随机数序列：

3179，1778，6301，2740，4975，3340，0247，9371，9769，0875，……

前10个随机数中，只有第7个、第10个分别为247、875，因小于1250可直接使用。其他8个随机数均需要取除以1250的余数，使其数值在1至1250之间。经过处理后得到：

679，528，51，240，1225，840，247，621，1019，875，……

由这些随机数作为编号的100个数据就是一个随机子样。实际抽样时，都是由计算机完成。实施步骤是将容量为 N 的一组数据存入一个数据文件，计算时先将文件读入一维数组。由介于1到 N 之间的 M 个随机数作为下标的数组元素值，就是一个容量为 M 的随机子样。

二、系统取样

系统取样是按一定顺序，机械地每隔若干个单位抽取一个样品的方法。例如，在钻井过程中，按钻进深度每增加1 m取一包岩屑就是典型的系统取样方法。这种取样方法简单易行，但有时可能产生系统误差。由于总体的性质不同，被取样的个体间隔不同，其抽样误差也不相同。

三、分层取样

分层取样是先按某种地质特征把研究对象分为若干个类型、部分或区域等，可通称分为若干个层。例如，按不同岩石类型取样，按不同勘探目的层取样，不同盆地进行取样等等都是分层

取样。

分层取样可按随机取样方式进行，如果在各层内抽样比例相同，可称按比例分层取样。当然，也可以不按比例进行取样。分层取样的目的是要把总体分为若干个部分。因此，一般情况下要求每个层的内部差异越小越好，而层与层之间的差异越大越好。

四、群体取样

群体取样不是抽取单个个体，而是抽取由个体组成的若干个集团，即个体群。例如在研究古生态环境时，取单一古生物化石往往不能说明问题，而用由各种古生物个体组成的古生物群落更反映古生态环境。

群体取样要求群体中包含的个体类型越多越好，以有利于说明其群体特征。

§4 地质变量

一、地质变量的概念

在地质研究过程中，一些地质标志或特征在不同的时间或空间内是不断变化的，如生油岩的厚度、埋藏深度，有机质的类型和丰度、成熟度等都随空间或时间的不同而发生变化。为了表示这些地质标志或特征的变化情况，就需要用变量来表示它们，这些变量就是地质变量。一般将地质变量定义为：反映地质系统中各成分标志或特征在时间和空间上变化情况的变量。

二、地质变量的分类

当地质变量取为某一个常数时，该常数就是地质数据，根据地质数据的不同，一般可将地质变量分为二种不同的类型。

（一）观测变量

地质变量的取值是通过对地质标志或特征直接进行观测、分析或度量所获得的各种原始观测值。观测变量是数学地质中最常见的一类变量，而其观测值是地质研究过程中的基本数据对象。

（二）综合变量

将二个或二个以上观测变量的观测值按特定的方式综合，形成新的综合性数据，该数据可视为综合地质变量的取值。因此，综合地质变量就是将二个或二个以上观测变量按特定的方式综合后形成的新地质变量，它具有特定的地质意义。怎样形成有意义的综合变量？这是具有丰富工作经验或具有灵感的地质人员通过假设、分析研究等手段，并经过实际验证后确定的。综合变量往往能够提供某些新的、重要的和隐蔽的地质信息，还可以起到减少观测变量、简化数学模型的作用。例如：

1. 用于区分天然气成因类型的甲烷系数 M

$$M = C_1/\sum_{i=1}^{5} C_i \tag{2-20}$$

当 M 小于等于 99%时，认为是热解成因气；否则认为是生物成因气。

2. 用于描述有机质成熟度的时间—温度指数 TTI(N. V. Lopatin, 1971)

$$TTI = \sum_{k=\min}^{\max} r^k(\Delta t) \tag{2-21}$$

一般认为：$TTI=15$，表示生油开始。$TTI=75$，表示生油高峰。$TTI=160$，表示生油结束。

三、地质变量的特征

地质观测工作是地质研究的基础，观测的结果就是各种地质资料，因此地质资料中包括了大量的基本地质信息。对于特定的地质研究对象而言，不是所有类型的地质信息都能成为有效的地质变量，作为地质变量，必须具有一定的特点。

1. *具有明确的地质意义*

地质变量的地质意义主要是指地质变量和特定地质研究对象的何种特征有关，就石油地质的研究范畴而言，一般包括以下几个方面的含义：

① 对地质变量所代表的地质特征的认识。如沉积盆地内生油岩的时代、历史上的最大埋深，地温梯度，圈闭的闭合面积、闭合高度等。

② 对地质变量所代表的地球化学特征的认识。如有机质的类型、丰度、干酪根的生油潜量、*TTI* 值、*OEP* 值、频率因子、活化能分布、油气化探指标异常等。

③ 对地质变量所代表的地球物理特征的认识。如盐丘通常为具有较大体积的岩盐穹形隆起或刺穿构造，其翼部较陡，顶部常接近于地面，岩盐具有明显小于围岩并且较均衡的密度，而覆盖盐丘的岩层通常又比较均匀，对重力观测结果的畸变影响不大。由于这些特征导致盐丘在重力图上显示为明显的负异常。图 2-9 为模拟盐丘处的计算重力剖面。

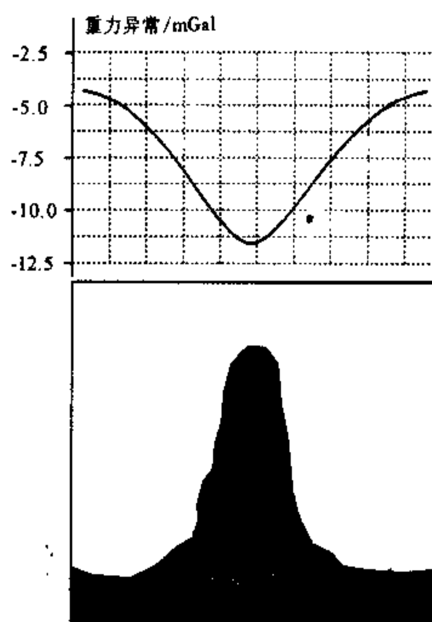


图 2-9 模拟盐丘处的重力计算剖面

④ 对地质变量所代表的其它方面特征的认识。如遥感地质测量、渗流力学等学科方面的特征和标志等。如地层异常压力和流体势等。

2. *具有明显的统计特征*

一般认为地质变量具有随机性和确定性的双重特征。对于随机性地质变量来说，研究它的统计分布特征往往可以从某个角度去解释地质现象和揭示地质规律。如有效的油气地表化探指标往往可以反映地下油气的分布规律性。一般，地质变量的统计特征越明显，它所反映的地质规律性就越强。因此，明显的统计特征是地质变量应具备的重要特征。

3. *地质变量与研究对象之间具有密切的关系*

地质变量与研究对象之间的关系越密切，该地质变量反映地质规律的能力就越强。如碎屑岩储集层中流体的饱和度与有效渗透率有关，其中之一发生变化时，另一个也发生变化。但这种变化关系因岩性不同而不同。试图建立饱和度与有效渗透率间的数学关系模型时，需要考虑到影响二者关系的主要因素，如粘土的膨胀作用、吸附膜、抗水表面及亲水表面、非混合性的其它流体以及气体压力等，如果以这些因素为变量，来搞清它们与有效渗透率和饱和度之间的数量关系，最终就有可能建立能描述饱和度与渗透率之间定量关系的数学模型，这些因素就成为研究上述定量关系的有效地质变量。对那些与研究对象之间数量关系不明确的地质信息，一般不能作为有效的地质变量来使用。

习 题

1. 简述地质数据的概念及分类。
2. 地质数据有什么特点？
3. 简述地质数据的误差类型。
4. 描述地质数据矩阵的一般形式。
5. 什么是地质数据的预处理？为什么要进行预处理？
6. 简述对地质数据标准化常用方法的变换公式和变换后的数据特点。
7. 怎样将定性数据转化为定量数据？
8. 叙述对原始数据进行网格化、减缩和增补的一般方法。
9. 什么是离群数据？怎样挑选和处理离群数据？
10. 简述对地质数据的取样方法。
11. 什么是地质变量？地质变量有哪几种？地质变量应该具备什么特征？

第三章　回归分析

§1　回归分析的概念及解决的问题

一、回归分析的概念

如果 y、$x_i(i=1,2,\cdots,m)$是描述 $m+1$ 个事物内在联系的变量，那么 y 与 $x_i(i=1,2,\cdots,m)$之间的关系大致可分为 2 种类型：

确定型关系，即函数关系。如，质点沿 x 轴从坐标 x_0 开始以初速度 v_0 和加速度 a 作匀加速直线运动，在任一时刻 t 质点的位置 x 由式

$$x=x_0+v_0t+\frac{1}{2}at^2$$

确定。又如，曲边梯形的面积 s 是曲边纵坐标 $f(x)$在其底边区间$[a,b]$上的定积分，即

$$s=\int_a^b f(x)\mathrm{d}x$$

一般情况下，把变量 y 与 $x_i(i=1,2,\cdots,m)$之间的函数关系记为

$$y=f(x_1,x_2,\cdots,x_m)$$

这种具有函数关系的变量是数学分析的研究对象。

不确定型关系，即相关关系。在地质学中，变量之间的关系是比较复杂的，往往是一个地质变量受其他一个或多个地质变量的制约，相互之间存在着一定程度的依赖性，但是，却又写不出它们之间所遵循的函数关系。地质学研究领域中诸如此类的问题很多，如有机质随矿物质沉积后演化为石油所需要的时间 t，主要是依赖于地层的温度 T，另外它也与地层的压力 P、有机质的性质及其他地质因素有关。一般说来，热演化温度高，有机质成熟快，演化为石油所需要的时间就短，反之就长。又如，一个含油气地质单元中石油资源量 Q 将随着含油气地质单元内生油岩的体积 V_1、储集岩的体积 V_2、近油源圈闭面积 S 的增大和有机质转化率 k 的升高而增多，却随着含油气地质单元内经受剥蚀次数 n 的增多而减少。上述这类地质变量的共同特点是某个地质变量 y 对另外 $m(m\geqslant 1)$个地质变量 $x_i(i=1,2,\cdots,m)$存在着一定程度的依赖性，但他们之间的数量关系却是不确定的，即不能由 $x_i(i=1,2,\cdots,m)$直接推测出 y 的值。在此，把这种数量关系不确定的变量称为相关变量，它是回归分析的研究对象。

设变量 y、$x_i(i=1,2,\cdots,m)$是相关变量，并在实际工作中获得了他们的 n 组观测值，记为

$$(x_{1k},x_{2k},\cdots,x_{mk},y_k)\qquad(k=1,2,\cdots,n)\tag{3-1}$$

回归分析就是通过式(3-1)给出的 n 组观测值研究变量间相关关系的一种统计分析方法，或者说，它是根据式(3-1)中的观测值确定相关变量间数学表达式的一种统计分析方法。

根据 m 的大小，回归分析可分为一元($m=1$)和多元($m\geqslant 2$)回归分析；按照变量间的相关类型，它可分为线性和非线性回归分析；从计算方法上，它又可分为"逐步剔除"、"逐步引入"、逐步回归和加权回归分析。

二、回归分析解决的问题

在地质研究工作中，回归分析主要解决三个方面的问题：其一是确定地质变量 y 与

$x_i(i=1,2,\cdots,P;P\leqslant m)$之间是否存在相关关系，如果存在，可以找出表示它们之间相关关系的数学表达式；其二是根据 $x_i(i=1,2,\cdots,p;p\leqslant m)$的观测值，利用确定出的数学表达式预测 y 的估计值，并且可以知道预测结果的精确度；其三是通过回归分析可以确定哪些地质变量对 y 的作用大，哪些变量对 y 的影响是无足轻重的，进而化简地质研究。

§2　多元线性回归分析

一、线性回归模型

若变量 y 与 $x_i(i=1,2,\cdots,m)$之间具有关系

$$y=a_0+\sum_{i=1}^{m}a_ix_i+\varepsilon \tag{3-2}$$

就说变量 y 与 $x_i(i=1,2,\cdots,m)$之间具有 m 元线性相关关系，简称为 m 元线性关系，并称式(3-2)为 m 元线性回归模型，其中 $a_0,a_1,\cdots,a_m$ 为待定系数，ε 为误差项，且 $\varepsilon—N(0,\sigma^2)$。在某个统计原则下，根据式(3-1)给出的 n 组观测值可以确定式(3-2)中待定系数的最佳估计值 b_0，$b_1,\cdots,b_m$，由此得一个方程，记为

$$\hat{y}=b_0+\sum_{i=1}^{m}b_ix_i \tag{3-3}$$

称式(3-3)为 $x_i(i=1,2,\cdots,m)$对 y 的线性回归方程，其中的 $b_0,b_1,\cdots,b_m$ 叫做回归系数。

二、确定回归系数

从数学上知，式(3-3)是 $m+1$ 维空间中的一个平面，如图 3-1 所示。对于任意一个观测点(x_{ik},y_k)，$(i=1,2,\cdots,m,k=1,2,\cdots,n)$在回归平面上有一个投影点$(x_{ik},\hat{y}_k)$，称观测值与回归值之差 $\delta_k=(y_k-\hat{y}_k)$为偏差(残差或剩余)。确定回归系数的统计原则是使 n 个偏差的平方和

$$Q_1=\sum_{k=1}^{n}\delta_k^2=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

达到最小。Q_1 是关于 $b_0,b_1,\cdots,b_m$ 的二次函数，且 $Q_1>0$，因此有

$$\begin{cases}\dfrac{\partial Q_1}{\partial b_0}=0\\[2mm]\dfrac{\partial Q_1}{\partial b_j}=0\end{cases}\qquad(j=1,2,\cdots,m)$$

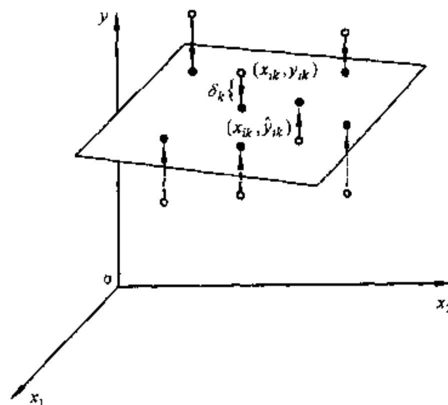


图 3-1　观测值与回归值偏差示意图

即

$$\begin{cases}\sum\limits_{k=1}^{n}(y_k-b_0-\sum\limits_{i=1}^{m}b_ix_{ik})=0\\[2mm]\sum\limits_{k=1}^{n}(y_k-b_0-\sum\limits_{i=1}^{m}b_ix_{ik})x_{jk}=0\end{cases}\qquad(j=1,2,\cdots,m) \tag{3-4}$$

由线性方程组(3-4)可解出 $b_0,b_1,\cdots,b_m$，得回归方程式(3-3)。

为了便于计算，改写式(3-4)的形式。

从式(3-4)的第一个方程解出

$$b_0 = \overline{y} - \sum_{i=1}^{m} b_i \overline{x}_i \tag{3-5}$$

把式(3-5)代入式(3-4)的后 m 个方程,并由式(3-4)的第一个方程等于零可得

$$\sum_{k=1}^{n}[(y_k - \overline{y}) - \sum_{i=1}^{m} b_i(x_{ik} - \overline{x}_i)](x_{jk} - \overline{x}_j) = 0 \tag{3-6}$$

$$(j = 1,2,\cdots,m)$$

把式(3-6)展开整理有

$$\sum_{k=1}^{n}(y_k - \overline{y})(x_{jk} - \overline{x}_j) = \sum_{i=1}^{m} b_i \sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j) \tag{3-7}$$

$$(j=1,2,\cdots,m)$$

令式(3-7)中

$$S_{jy} = \sum_{k=1}^{n}(y_k - \overline{y})(x_{jk} - \overline{x}_j) \qquad (j = 1,2,\cdots,m)$$

$$S_{ij} = \sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j) \qquad (i,j = 1,2,\cdots,m)$$

至此,可以把式(3-4)改写为

$$\begin{cases} b_0 = \overline{y} - \sum\limits_{i=1}^{m} b_i \overline{x}_i \\ \sum\limits_{j=1}^{m} S_{ij} b_j = s_{iy} \end{cases} \qquad (i = 1,2,\cdots,m) \tag{3-8}$$

线性方程组式(3-8)称为式(3-4)的正规方程组。从式(3-8)的后 m 个方程解出 $b_1,b_2,\cdots,b_m$,再把它们代入式(3-8)的第一个方程求出 b_0,于是就得到回归方程式(3-3)。

上述使偏差平方和最小确定待定系数的方法叫最小二乘法。

三、回归检验

假定变量 y 与 $x_i(i=1,2,\cdots,m)$之间存在线性关系,根据 n 组观测值,用最小二乘法确定了回归系数,建立了 $x_i(i=1,2,\cdots,m)$对 y 的回归方程。现在要问:线性相关的假设正确与否?求得的回归方程代表性如何?各个地质变量在回归方程中的作用又是怎样?为了回答上述问题,就要从统计分析上对它们进行检验,这就是回归显著性检验。它包括对回归方程的显著性检验和对各地质变量的显著性检验。

(一)对回归方程的显著性检验

为了构造检验统计量,先分解变量 y 与其平均值 $\overline{y}$ 的总偏差平方和 Q:

$$\begin{aligned} Q &= \sum_{k=1}^{n}(y_k - \overline{y})^2 = \sum_{k=1}^{n}[(y_k - \hat{y}_k) + (\hat{y}_k - \overline{y})]^2 \\ &= \sum_{k=1}^{n}(y_k - \hat{y}_k)^2 + 2\sum_{k=1}^{n}(y_k - \hat{y}_k)(\hat{y}_k - \overline{y}) + \sum_{k=1}^{n}(\hat{y}_k - \overline{y})^2 \\ &= Q_1 + Q_2 \end{aligned} \tag{3-9}$$

式中 Q_1——偏差平方和,$Q_1 = \sum_{k=1}^{n}(y_k - \hat{y}_k)^2$;

Q_2——回归平方和,$Q_2 = \sum_{k=1}^{n}(\hat{y}_k - \overline{y})^2$;

Δ—— 交叉项，　　$\Delta = \sum_{k=1}^{n}(y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) = 0$

证明交叉项等于 0 如下：

把 $\bar{y} = b_0 + \sum_{i=1}^{m} b_i \bar{x}_i$、$\hat{y}_k = b_0 + \sum_{j=1}^{m} b_j x_{jk}$ 代入 Δ 得

$$\Delta = \sum_{k=1}^{n}(y_k - b_0 - \sum_{j=1}^{m} b_j x_{jk}) \cdot \sum_{j=1}^{m} b_j (x_{jk} - \bar{x}_j)$$

再把 $b_0 = \bar{y} - \sum_{i=1}^{m} b_i \bar{x}_i$ 代入 Δ 得

$$\begin{aligned}\Delta &= \sum_{j=1}^{m} b_j \sum_{k=1}^{n}(y_k - \bar{y})(x_{jk} - \bar{x}_j) - \sum_{j=1}^{m} b_j \sum_{i=1}^{m} b_i \sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j) \\ &= \sum_{j=1}^{m} b_j s_{jy} - \sum_{j=1}^{m} b_j \sum_{i=1}^{m} b_i s_{ij} = \sum_{j=1}^{m} b_j s_{jy} - \sum_{j=1}^{m} b_j s_{jy} = 0\end{aligned}$$

式(3-9)表明，变量 y 与其平均值 $\bar{y}$ 的总偏差平方和 Q 可分解为两部分。其中的一部分为 Q_1，它是观测值 y_k 与回归值 $\hat{y}_k$ 之间的偏差平方和，主要是线性回归模型引起的偏差。另一部分为 Q_2，它是回归值 $\hat{y}_k$ 与 y 的平均值 $\bar{y}$ 之差的平方和，反映了变量 $x_i(i=1,2,\cdots,m)$ 的变化对 y 引起的波动，称为回归平方和。

Q、Q_1、Q_2 的自由度分别为：

$$f_Q = n-1;\quad f_{Q_1} = n-m-1;\quad f_{Q_2} = m$$

且满足等式

$$f_Q = f_{Q_1} + f_{Q_2}$$

1. 复相关系数检验

由式(3-9)可知，Q_1 愈小，Q_2 就愈接近于 Q，说明变量 y 与 $x_i(i=1,2,\cdots,m)$ 的线性关系越密切，回归模型(3-2)的偏差就越小，即回归方程所代表的变化关系就越接近实际。由此可用比值 Q_2/Q 作为衡量回归方程显著性的一个指标。定义

$$R = (Q_2/Q)^{1/2}$$

为 y 与 $x_i(i=1,2,\cdots,m)$ 的复相关系数。R 愈接近 ± 1，回归方程的显著性越高；反之，当 R 愈接近于 0 时，y 与 $x_i(i=1,2,\cdots,m)$ 的相关性越差，由观测值求得的回归方程就没有实际意义。

2. F 分布检验

假设 H_0：变量 y 与 $x_i(i=1,2,\cdots,m)$ 之间没有回归模型式(3-2)给出的线性相关关系。

若假设 H_0 为真，那么 Q_1 就比较大，Q_2 比较小，比值 $r=Q_2/Q_1$ 就小。当 r 小于某个临界值时，接受原假设 H_0；否则，就否定原假设 H_0，即变量 y 与 $x_i(i=1,2,\cdots,m)$ 间有着密切的线性关系。

若假设 H_0 为真，可以证明统计量

$$F = \frac{Q_2/f_{Q_2}}{Q_1/f_{Q_1}} = \frac{Q_2/m}{Q_1/(n-m-1)} = \frac{Q_2 \cdot (n-m-1)}{m \cdot Q_1} \tag{3-10}$$

服从第一自由度为 m、第二自由度为 $(n-m-1)$ 的 $F(m,n-m-1)$ 分布。根据给定的检验水平 α，在 $F_\alpha(m,n-m-1)$ 分布表上查出临界值 F_α。当 $F > F_\alpha$ 时，否定原假设 H_0，这时称回归方程是显著的，可以付诸应用；若 $F \leqslant F_\alpha$，则肯定原假设 H_0，即求得的回归方程没有实际意义。

（二）对变量的显著性检验

不论是用复相关系数 R 还是用统计量 F 对回归方程进行显著性检验，都是检验的 m 个地质变量对 y 的综合作用。实际上，变量 $x_i(i=1,2,\cdots,m)$ 对 y 的作用是不一样的，甚至有的变量与 y 就没有线性关系。一般总希望回归方程中不包含那些无关紧要的变量，因此就要对拟定的变量 $x_i(i=1,2,\cdots,m)$ 逐个进行检验。如果变量 x_r 与 y 无线性相关关系，就相当于线性回归模型中 x_r 的待定系数 a_r 等于 0。因此，对变量 x_r 的显著性检验，就等于对如下的假设 H_0 进行检验。

假设 H_0：$a_r=0$

统计量

$$F_r=\frac{(b_r-a_r)^2/s_{rr}^{-1}}{Q_1/(n-m-1)}=\frac{b_r^2/s_{rr}^{-1}}{Q_1/(n-m-1)} \tag{3-11}$$

服从第一自由度为 1，第二自由度为 $(n-m-1)$ 的 $F(1,n-m-1)$ 分布。式(3-11)中的 s_{rr}^{-1} 是式(3-8)后 m 个方程系数矩阵 S 的逆矩阵 S^{-1} 中第 r 行第 r 列的元素。

对于给定的检验水平 α，由 $F_\alpha(1,n-m-1)$ 分布表查得临界值 F_α。若 $F_r>F_\alpha$，则在检验水平 α 下拒绝原假设 H_0，即认为变量 x_r 对 y 的作用显著。反之，若 $F_r\leqslant F_\alpha$，在检验水平 α 下接受原假设 H_0，变量 x_r 不应进入回归方程。

对变量 $x_i(i=1,2,\cdots,m)$ 逐个进行检验，筛选出对 y 作用显著的变量重新进行回归分析，建立更为简单而有效的线性回归方程。

四、回归预测与控制

在变量 y 与 $x_i(i=1,2,\cdots,m)$ n 组观测值

$$(x_{1k},x_{2k},\cdots,x_{mk},y_k) \qquad (k=1,2,\cdots,n)$$

的基础上，根据最小二乘法原理建立了 $x_i(i=1,2,\cdots,m)$ 对 y 的回归方程

$$\hat{y}=b_0+\sum_{i=1}^{m}b_ix_i$$

经过检验，如果回归方程以及各变量都是显著的，那么就可以利用这个回归方程对 y 进行预测或控制。

所谓预测，就是把给定的 $x_{ir}(i=1,2,\cdots,m)$ 代入回归方程算出

$$\hat{y}_r=b_0+\sum_{i=1}^{m}b_ix_{ir}$$

用 $\hat{y}_r$ 作为 y_r 的估计值。剩余标准差

$$\hat{\sigma}=\sqrt{Q_1/(n-m-1)}$$

刻划了 y_r 的取值范围及相应的取值概率。当变量 $x_i(i=1,2,\cdots,m)$ 取定值 $x_{ir}(i=1,2,\cdots,m)$ 并且 $\min\limits_{k=1,2,\cdots,n} x_{ik}\leqslant x_{ir}\leqslant \max\limits_{k=1,2,\cdots,n} x_{ik}$ 时，对应的 y_r 落在 $(\hat{y}_r-\hat{\sigma},\hat{y}_r+\hat{\sigma})$，$(\hat{y}_r-2\hat{\sigma},\hat{y}_r+2\hat{\sigma})$ 和 $(\hat{y}_r-3\hat{\sigma},\hat{y}_r+3\hat{\sigma})$ 内的概率大约分别为 68%、95% 和 99%。

所谓控制问题，就是调整 $x_i(i=1,2,\cdots,m)$，使 y 值落在某一个给定的范围 $y_1\leqslant y\leqslant y_2$ 内。当给定值 $x_{ir}(i=1,2,\cdots,m)$ 满足条件

$$(\hat{y}_r-2\hat{\sigma}\geqslant y_1); \quad (\hat{y}_r+2\hat{\sigma}\leqslant y_2)$$

时，即可使 y 在给定范围 $y_1\leqslant y\leqslant y_2$ 内取值。

五、非线性回归分析

在地质研究工作中，经常遇到呈现出非线性相关的地质变量，如，岩石的有效孔隙度 φ_e 与

总孔隙度 φ 之间的关系为：

$$\varphi_e = a_0 + a_1\varphi^2$$

又如，岩石渗透率 k 与声波时差 Δt、自然伽玛相对值 ΔGR 之间有

$$\ln k = a_0 + a_1 \ln\Delta t + a_2 \Delta GR$$

的相关关系。

对于表现为非线性相关的地质变量，一般先用变量替换的方法，把它们化为线性回归模型，然后再用线性回归分析方法求取回归方程。

六、多元线性回归分析的计算步骤

（一）关于计算某些量的简便公式

为了化简计算，先化简 Q、Q_1、Q_2 及 S_{ij} 计算公式，并且令 $y = x_{m+1}$。

$$Q = \sum_{k=1}^{n}(x_{m+1,k} - \bar{x}_{m+1})^2 = \sum_{k=1}^{n} x_{m+1,k}^2 - 2\bar{x}_{m+1}\sum_{k=1}^{n} x_{m+1,k} + \sum_{k=1}^{n}\bar{x}_{m+1}^2$$

$$= \sum_{k=1}^{n} x_{m+1,k}^2 - n\,\bar{x}_{m+1}^2 \tag{3-12}$$

$$Q_1 = Q - Q_2 \tag{3-13}$$

$$Q_2 = \sum_{k=1}^{n}(\hat{x}_{m+1,k} - \bar{x}_{m+1})^2 = \sum_{k=1}^{n}(b_0 + \sum_{i=1}^{m} b_i x_{ik} - \bar{x}_{m+1})^2$$

$$= \sum_{k=1}^{n}(\bar{x}_{m+1} - \sum_{i=1}^{m} b_i \bar{x}_i + \sum_{i=1}^{m} b_i x_{ik} - \bar{x}_{m+1})^2$$

$$= \sum_{k=1}^{n}\left[\sum_{i=1}^{m}\sum_{j=1}^{m} b_i b_j (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)\right] = \sum_{i=1}^{m} b_i s_{i,m+1} \tag{3-14}$$

$$s_{ij} = \sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j) = \sum_{k=1}^{n} x_{ik}\cdot x_{jk} - \sum_{k=1}^{n} x_{jk}\cdot\bar{x}_i$$

$$- \sum_{k=1}^{n} x_{ik}\cdot\bar{x}_j + \sum_{k=1}^{n}\bar{x}_i\cdot\bar{x}_j = \sum_{k=1}^{n} x_{ik}\cdot x_{jk} - n\bar{x}_i\bar{x}_j \tag{3-15}$$

$$(i = 1, 2, \cdots, m; \quad j = 1, 2, \cdots, m+1)$$

（二）计算步骤

多元线性回归分析的计算步骤如图 3-2 所示。

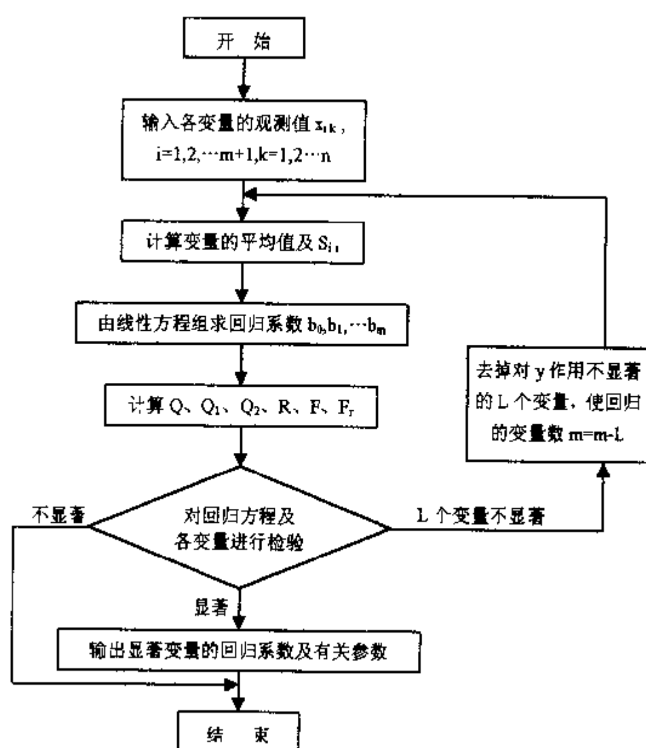


图 3-2　多元线性回归流程图

§3　逐步回归分析

一、逐步回归分析的由来

在地质研究工作中，依据经验或在某地质理论指导下，拟定了 m 个认为与变量 y 有着密切联系的变量 $x_i(i=1,2,\cdots,m)$，是否真是这样呢？也许 x_i 中有部分变量被认为对预测 y 起着重要的作用，但从它们的观测值分析却不一定符合人们的经验认识。变量 $x_i(i=1,2,\cdots,m)$ 对 y 作用的大小，可以通过对回归系数进行的假设检验来鉴别，正象多元线性回归分析中指出的那样，如果变量 x_i 的回归系数 $b_i=0$ 这个假设被接受，就反映变量 x_i 对 y 的作用不重要，这时就从回归方程中把它去掉，重新建立回归方程，这样可使不包含变量 x_i 的回归方程更合理地反映 y 与 $x_i(i=1,2,\cdots,i-1,i+1,\cdots,m)$ 的相关性。由此可以说，上一节中介绍的多元线性回归分析是一种“逐步剔除”变量、循序渐进地寻找变量间相关关系的回归分析方法。它的基本过程是：

第一步，先建立变量 $x_i(i=1,2,\cdots,m)$ 对 y 的回归方程，然后对回归系数 $b_i(i=1,2,\cdots,m)$ 逐个进行检验，剔除 $x_i(i=1,2,\cdots,m)$ 中使假设 $H_0:b_i=0$ 被接受的变量。为叙述方便，无妨设保留下来是 $x_i(i=1,2,\cdots,m)$ 中的前 $P(P<m)$ 个变量。

第二步，重新建立保留下来的变量 $x_i(i=1,2,\cdots,p)$ 对 y 的回归方程，并逐个检验回归系数 $b_i(i=1,2,\cdots,p)$，看是否仍有使假设 $H_0:b_i=0$ 被接受的变量。如果有就从变量 $x_i(i=1,2,\cdots,p)$ 中剔除相应的变量，重复上述过程，总会出现使假设 $H_0:b_i=0$ 全部被拒绝的一步，这时

就获得最终的回归方程。

上述"逐步剔除"变量建立回归方程的方法有一个明显的缺点，就是把每一步保留下来的变量全部引入回归方程，再逐个检验是否剔除，计算工作量很大，实际上有些不重要的变量就不必引入。基于这种考虑，就又提出了"逐步引入"的回归分析方法。它的基本过程如下：

第一步，先分别建立 $x_i(i=1,2,\cdots,m)$ 对 y 的回归方程，记为

$$\hat{y}_i=b_{0i}+b_{1i}x_i \qquad (i=1,2,\cdots,m)$$

然后逐个检验 $\hat{y}_i$ 的显著性，从中选出一个最显著的回归方程 $\hat{y}_r$，并且把与 $\hat{y}_r$ 相应的变量 x_r 引入回归方程，无妨设 x_r 是 x_1。

第二步，逐个比较包含 $(x_1,x_2),(x_1,x_3),\cdots,(x_1,x_m)$ 的回归方程，检验方程中增加变量 x_i $(i\neq1)$ 之后，x_i 的回归系数是否显著地不为 0，再在显著不为 0 的回归方程中选出一个最显著的方程 $\hat{y}'_r$，并把与 $\hat{y}'_r$ 相应的变量 x_r 再引入回归方程，无妨设第二步引入的 x_r 为 x_2。

第三步，在回归方程中已引入 x_1,x_2 的基础上，再逐个添加变量 $x_i(i=3,4,\cdots,m)$，并检验引入新变量后的回归方程中是否有比包含变量 x_1,x_2 的回归方程有显著改进的回归方程，有就再引入新的变量，…这样重复进行，直到没有可被引入的变量为止，这时就建立了最终的回归方程。

这种"逐步引入"变量的回归分析方法，没有考虑到 $x_i(i=1,2,\cdots,m)$ 间的相关性，也就是说，在逐步引入的过程中，后引入的变量有可能使先引入的变量由对 y 作用重要而变为不显著，当出现这种情况时，就应及时从回归方程中去掉作用不重要的变量。然而，"逐步引入"回归法却不能实现这一要求。

吸收上述两种多元线性回归方法的优点，就派生出了边"引入"边"剔除"变量的逐步回归分析。它的基本思想是：在回归分析的过程中，根据变量 $x_i(i=1,2,\cdots,m)$ 对 y 作用的大小，依次引入到回归方程中，同时还要对引入回归方程中的变量逐个检验，及时剔除其中对 y 作用不显著的变量。照此进行下去，直到没有对 y 作用显著的变量可引入回归方程，同时方程中也没有对 y 作用不显著的变量被剔除为止，这时的回归方程中就仅包含了对 y 作用显著的变量。

由上述可知，逐步回归分析的优点是能够从数量较多的变量中筛选出对 y 作用重要的变量引入回归方程，从而克服了"逐步剔除"和"逐步引入"回归分析方法存在的不足。另外，在不知道变量相关形式的情况下，同一个变量以不同形式出现时，如 $x,x^2,\ln x,\sin x$ 等，可以看作是不同的变量，利用逐步回归分析能够筛选变量的特点，又可以帮助我们确定变量间的相关形式。

究竟怎样选出对 y 作用重要的变量呢？下面就介绍"引入"和"剔除"变量的原则。

二、"引入"和"剔除"变量的原则

1. "引入"变量的原则

假设回归方程中已引入了 l 个变量 $x_{k_1},x_{k_2}\cdots,x_{k_l}$，对应的回归方程是

$$\hat{y}=b_0+b_{k_1}x_{k_1}+\cdots+b_{k_l}x_{k_l}$$

它的总偏差平方和 Q 的分解式是

$$Q=Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l})+Q_2(x_{k_1},x_{k_2}\cdots,x_{k_l})$$

在回归方程中再增加一个变量 $x_{k_i}(k_i\overline{\in}(k_1,k_2,\cdots,k_l)$ 后，就得到含有 $(l+1)$ 个变量的回归方程

$$\hat{y}=\hat{b}_0+\hat{b}_{k_1}x_{k_1}+\cdots+\hat{b}_{k_l}x_{k_l}+\hat{b}_{k_i}x_{k_i}$$

相应的总偏差平方和 Q 的分解式就是

$$Q=Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_i})+Q_2(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_i})$$

比较回归方程中有 l 和$(l+1)$个变量时的总偏差平方和 Q 的分解式，有

$$Q_2(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_i})-Q_2(x_{k_1},x_{k_2}\cdots,x_{k_l})$$

$$=Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l})-Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_i})$$

记 $V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})=Q_2(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_i})-Q_2(x_{k_1},x_{k_2}\cdots,x_{k_l})$或

$$V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})=Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l})-Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_i})$$

$V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})$是变量 x_{k_i}引入后，对 y 引起的波动，它即是回归平方和的增加量，又是剩余平方和的减少量，称 $V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})$是变量 x_{k_i}对 y 的方差贡献。当 x_{k_i}与 y 的相关程度较低时，方差贡献 $V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})$就比较小，而剩余平方和 $Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_i})$就比较大。那么，计算出一切不在回归方程内的$(m-l)$个变量的方差贡献，选出其中最大者，记为 V_{k_a}，即

$$V_{k_a}=\max_{\substack{1\leqslant k_i\leqslant m\\ k_i\neq k_1,k_2,\cdots,k_l}}V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})$$

和 V_{k_a}对应的变量是 x_{k_a}。根据上述讨论，可以提出如下假设 H_0：

假设 H_0：变量 x_{k_a}与 y"无"线性相关关系。

统计量

$$F_{k_a}=\frac{V_{k_a}(x_{k_1},x_{k_2}\cdots,x_{k_l})/1}{Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l},x_{k_a})/(n-l-2)} \tag{3-16}$$

遵从 $F(1,n-l-2)$分布。

式(3-16)中 n 为样本容量(数据组数)，l 是已"引入"的变量数。

适当地选取引入变量的临界值 F_1，当

$$F_{k_a}>F_1$$

时，否定原假设 H_0，即变量 x_{k_a}对 y 的影响大，应该把变量 x_{k_a}引入回归方程。当

$$F_{k_a}\leqslant F_1$$

时，接收原假设 H_0，即变量 x_{k_a}与 y 无线性相关关系，则变量 x_{k_a}不能引入回归方程，引入变量结束(因为方差贡献最大的变量 x_{k_a}都没有资格进入回归方程，其余方差贡献小的变量就更不够条件了)。

2. "剔除"变量的原则

假定回归方程中已引入了 $x_{k_1},x_{k_2}\cdots,x_{k_l}$共 l 个变量，这时的回归方程是

$$\hat{y}=b_0+b_{k_1}x_{k_1}+\cdots+b_{k_l}x_{k_l}$$

该回归方程的总偏差平方和 Q 的分解式是

$$Q=Q_1(x_{k_1},x_{k_2}\cdots,x_{k_l})+Q_2(x_{k_1},x_{k_2}\cdots,x_{k_l})$$

现在逐个检查已进入回归方程的 l 个变量 $x_{k_1},x_{k_2},\cdots,x_{k_l}$，把对 y 作用变得不显著的变量从回归方程中剔除出去。

设变量 x_{k_i}($k_i\in(k_1,k_2,\cdots,k_l)$被剔除，剔除变量 x_{k_i}后，回归方程的总偏差平方和 Q 的分解式是

$$Q=Q_1(x_{k_1},x_{k_2}\cdots,x_{k_{i-1}},x_{k_{i+1}},\cdots,x_{k_l})+Q_2(x_{k_1},x_{k_2},\cdots,x_{k_{i-1}},x_{k_{i+1}},\cdots,x_{k_l})$$

比较回归方程中有 l 和$(l-1)$个变量时的总偏差平方和 Q 的分解式，有

$$Q_2(x_{k_1},x_{k_2},\cdots,x_{k_l})-Q_2(x_{k_1},x_{k_2},\cdots,x_{k_{i-1}},x_{k_{i+1}},\cdots,x_{k_l})$$

$$=Q_1(x_{k_1},x_{k_2},\cdots,x_{k_{i-1}},x_{k_{i+1}},\cdots,x_{k_l})-Q_1(x_{k_1},x_{k_2},\cdots,x_{k_l})$$

记 $V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})=Q_1(x_{k_1},x_{k_2}\cdots,x_{k_{i-1}},x_{k_{i+1}},\cdots,x_{k_l})-Q_1(x_{k_1},x_{k_2},\cdots,x_{k_l})$

$V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})$是变量 x_{k_i}被剔除后回归平方和的减少量，又是剩余平方和的增加量。由此可知，$V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})$是变量 x_{k_i}的方差贡献。算出回归方程中 l 个变量 $x_{k_1},x_{k_2},\cdots,x_{k_l}$对 y 的方差贡献，从中选出最小的，记为 V_{k_a}，即

$$V_{k_a}=\min_{i=1,2,\cdots,l}V_{k_i}(x_{k_1},x_{k_2}\cdots,x_{k_l})$$

和引入变量一样，提出如下假设：

假设 H_0：变量 x_{k_a}对 y"作用不显著"

统计量

$$F'_{k_a}=\frac{V_{k_a}(x_{k_1},x_{k_2},\cdots,x_{k_l})/1}{Q_1(x_{k_1},\cdots,x_{k_l})/(n-l-1)} \qquad (3\text{-}17)$$

遵从 $F(1,n-l-2)$分布。

F'_{k_a}愈小，变量 x_{k_a}在回归方程中起的作用就愈不显著。适当选取剔除变量的临界值 F_2，

$$F'_{k_a}\leqslant F_2$$

时，首先从回归方程中剔除变量 x_{k_a}，然后考虑回归方程中是否还有要剔除的变量，直到回归方程中没有对 y 影响不显著的变量剔除时，就可转入是否还有新变量引入的问题。当

$$F'_{k_a}>F_2$$

时，变量 x_{k_a}应保留在回归方程中，即回归方程中无任何不显著的变量可剔除（因为方差贡献最小的变量 x_{k_a}都不能被剔除，则方差贡献大的就不用考虑了）。

在 $F(1,n-l-2)$和 $F(1,n-l-1)$中，$n\gg l$，所以有

$$F_1\approx F_2$$

故可取 $F_1=F_2=F^*$，据经验取 $F^*=1,2,3,4$ 即可。

三、逐步回归的变换公式

逐步回归是在多元线性回归基础上派生的计算技巧，它是通过对变量的相关系数增广矩阵施实一系列矩阵变换来实现逐步引入或剔除变量，求解回归方程。

1. 相关系数增广矩阵

当由于变量的量纲不同，而导致观测值的数量级相差很大时，在回归分析中就存在两个方面的问题：其一是变量的量纲不同，无法在同一个数量级的基础上比较 $x_i(i=1,2,\cdots,m)$对 x_{m+1}作用的大小；其二是在数量级差异甚大的情况下，由式(3-15)所求得的线性方程组式(3-8)的系数及常数项，就会出现更大的差别，由此将给线性方程组的解带来较大的舍入误差，甚至使方程组无解。基于上述考虑，需要对变量观测值进行变换处理。在第二章里介绍了多种数据预处理方法，为适合逐步回归算法，这里采用标准差标准化把变量观测值处理成平均值等于0方差为1的无量纲数据，并把处理后得到的新变量称为标准化变量，记为 $x'_i(i=1,2,\cdots,m)$。

根据式(3-8)，标准化变量的回归系数 $b'_j(j=1,2,\cdots,m)$应满足：

$$\sum_{j=1}^{m}r_{ij}b'_j=r_{im+1} \qquad (i=1,2,\cdots,m) \qquad (3\text{-}18)$$

由式(3-15)和标准差标准化可知：

$$r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} \qquad (i,j=1,2,\cdots,m) \tag{3-19}$$

式(3-19)表明，式(3-18)的系数及常数项是原变量的相关系数。

方程组(3-18)的系数矩阵增加一行一列，得矩阵

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} & r_{1\,m+1} \\ r_{21} & r_{22} & \cdots & r_{2m} & r_{2\,m+1} \\ \vdots & & & & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} & r_{m\,m+1} \\ r_{m+1\,1} & r_{m+1\,2} & \cdots & r_{m+1\,m} & r_{m+1\,m+1} \end{bmatrix} \tag{3-20}$$

矩阵 $R=[r_{ij}](i,j=1,2,\cdots,m+1)$叫做相关系数增广矩阵。

2. 变换公式

为了区别变换前后的矩阵，记 $R^{(0)}$是初始的相关系数增广矩阵，$R^{(N)}$是 $R^{(0)}$经过 N 次变换后的矩阵，并叫它是相关增广矩阵。逐步回归是在回归方程中含有 l 个变量的基础上，逐步地引入或剔除变量。

设逐步回归进行了 N 步，此时回归方程中已包含 l 个变量 $x'_{k_1},x'_{k_2}\cdots,x'_{k_l}$。逐步回归的第 $N+1$ 步是在第 N 步基础上引入或剔除变量 x'_{k_a}。$R^{(N)}$的元素按公式(3-21)进行一次变换，得到第 $N+1$ 步的相关增广矩阵 $R^{(N+1)}$。

$$r_{ij}^{(N+1)}=\begin{cases} r_{ij}^{(N)}/r_{k_ak_a}^{(N)} & i=k_a,j\neq k_a; \\ r_{ij}^{(N)}-r_{ik_a}^{(N)}r_{k_aj}^{(N)}/r_{k_ak_a}^{(N)} & i\neq k_a,j\neq k_a; \\ -r_{ij}^{(N)}/r_{k_ak_a}^{(N)} & i\neq k_a,j=k_a; \\ 1/r_{k_ak_a}^{(N)} & i=k_a,j=k_a; \end{cases} \tag{3-21}$$

变换公式(3-21)来源于矩阵变换解线性方程组，有关《线性代数计算方法》书上有详细讨论，这里不再重述。变换公式有以下主要性质：

(1) 对称性

$$r_{ij}^{(N+1)}=\begin{cases} r_{ji}^{(N+1)}, & \text{当 } i、j\in(k_1,k_2,\cdots,k_l)\text{ 或 } i,j\overline{\in}(k_1,k_2,\cdots,k_l)\text{ 时} \\ -r_{ji}^{(N+1)}, & \text{当 } i\in(k_1,k_2,\cdots,k_l)\text{ 而 } j\overline{\in}(k_1,k_2,\cdots,k_l)\text{ 或} \\ & \text{当 } j\in(k_1,k_2,\cdots,k_l)\text{ 而 } i\overline{\in}(k_1,k_2,\cdots,k_l)\text{ 时} \end{cases}$$

对称性说明，在 $R^{(N+1)}$中，位于包含在 $N+1$ 步回归方程中变量位置上的元素及位于一切未被选入 $N+1$ 步回归方程变量位置上的元素都具对称性；$R^{(N+1)}$中，除上述位置外的元素则具反对称性。在逐步回归中，就利用这一性质判断变量 x'_{k_a}是否在 $N+1$ 步的回归方程中；另外，对称性还告诉我们，在计算 $R^{(N+1)}$时，只要计算出它的主对角线及其右上角的元素就够了。

(2) 自反性

所谓自反性，就是对同一个变量施行两次变换，则矩阵元素不变，即 $R^{(N+1)}=R^{(N-1)}$。

由自反性知，第 $N+1$ 步不论是引入还是剔除变量 x'_{k_a}都是按公式(3-21)对前一步矩阵 $R^{(N)}$施行一次变换，不同之处仅是变量 x'_{k_a}是否在第 N 步回归方程中。若第 N 步回归方程不包含变量 x'_{k_a}，施行第 $N+1$ 步变换的结果是把变量 x'_{k_a}引入了 $N+1$ 步回归方程，当变量 x'_{k_a}在第 N 步回归方程中时，第 $N+1$ 步变换就从第 N 步回归方程中剔除了变量 x'_{k_a}。

(3) $N+1$ 步矩阵 $R^{(N+1)}$只与 $N+1$ 步变量的全体有关，而与变量引入的先后次序及曾经

引入后又被剔除的变量无关。

四、剩余平方和、方差贡献及 F 统计量的计算

在逐步回归计算过程中，要从 $x'_i(i=1,2,\cdots,m)$ 中选出对 y' 方差贡献大而又显著的变量引入回归方程，同时又要从已进入回归方程的变量中找出对 y' 方差贡献小而又不显著的变量剔除，就要不断计算变量的方差贡献，下面导出计算方差贡献的公式。

假设 $x'_{k_1},x'_{k_2},\cdots,x'_{k_l}$ 是 m 个变量中对 y' 的变化起着重要作用的 l 个变量，则与 $x'_{k_1},x'_{k_2},\cdots,x'_{k_l}$ 对应的方程组

$$\sum_{j=1}^{l} r_{k_ik_j}b'_{k_j}=r_{k_i m+1} \qquad (i=1,2,\cdots,l) \tag{3-22}$$

的相关系数增广矩阵

$$\widetilde{R}=\begin{bmatrix} r_{k_1k_1} & r_{k_1k_2} & \cdots & r_{k_1k_l} & r_{k_1m+1} \\ r_{k_2k_1} & r_{k_2k_2} & \cdots & r_{k_2k_l} & r_{k_2m+1} \\ \vdots & & & & \vdots \\ r_{k_lk_1} & r_{k_lk_2} & \cdots & r_{k_lk_l} & r_{k_lm+1} \\ r_{m+1\,k_1} & r_{m+1\,k_2} & \cdots & r_{m+1\,k_l} & r_{m+1\,m+1} \end{bmatrix}$$

在 $\widetilde{R}_{m+1\,m+1}\neq 0$ 的条件下，方程组(3-22)的解是

$$b'_{k_i}=-\frac{\widetilde{R}_{m+1\,k_i}}{\widetilde{R}_{m+1\,m+1}} \qquad (i=1,2,\cdots,l) \tag{3-23}$$

其中 $\widetilde{R}_{m+1\,k_i}$ 是 $\widetilde{R}$ 中 $r_{m+1\,k_i}$ 的代数余子式，即

$$\widetilde{R}_{m+1\,k_i}=(-1)^{m+1+k_i}\begin{vmatrix} r_{k_1k_1} & \cdots & r_{k_1k_i-1} & r_{k_1k_i+1} & \cdots & r_{k_1m+1} \\ r_{k_2k_1} & \cdots & r_{k_2k_i-1} & r_{k_2k_i+1} & \cdots & r_{k_2m+1} \\ \vdots & & & & & \vdots \\ r_{k_lk_1} & \cdots & r_{k_lk_i-1} & r_{k_lk_i+1} & \cdots & r_{k_lm+1} \end{vmatrix}$$

$$\widetilde{R}_{m+1\,m+1}=\begin{vmatrix} r_{k_1k_1} & r_{k_1k_2} & \cdots & r_{k_1k_l} \\ r_{k_2k_1} & r_{k_2k_2} & \cdots & r_{k_2k_l} \\ \vdots & & & \vdots \\ r_{k_lk_1} & r_{k_lk_2} & \cdots & r_{k_lk_l} \end{vmatrix}$$

若 $\widetilde{R}$ 用普通消元法就可化为对角矩阵

$$\begin{bmatrix} r_{k_1k_1} & & & & 0 \\ & r'_{k_2k_2} & & & \\ & & r''_{k_3k_3} & & \\ & & & \ddots & \\ 0 & & & & r'^{(l)}_{m+1\,m+1} \end{bmatrix}$$

则

$$|\widetilde{R}|=r_{k_1k_1}r'_{k_2k_2}r''_{k_3k_3}\cdots r'^{(l)}_{m+1\,m+1} \tag{3-24}$$

同理有

$$\widetilde{R}_{m+1\,m+1}=r_{k_1k_1}r'_{k_2k_2}r''_{k_3k_3}\cdots r'^{(l-1)}_{k_lk_l} \tag{3-25}$$

由(3-24)、(3-25)得

$$r'^{(l)}_{m+1\,m+1}=|\tilde{R}|/\tilde{R}_{m+1\,m+1} \tag{3-26}$$

矩阵变换法是消去法的一种变形，但逐步回归中的矩阵变换不必作行列交换；另外我们再规定最后一个主对角元不参加选主元，那么相关系数增广矩阵经 N 步矩阵变换后得 $R^{(N)}$

$$R^{(N)}=\begin{bmatrix} r_{11}^{(N)} & r_{12}^{(N)} & \cdots & r_{1m}^{(N)} & r_{1\,m+1}^{(N)} \\ r_{21}^{(N)} & r_{22}^{(N)} & \cdots & r_{2m}^{(N)} & r_{2\,m+1}^{(N)} \\ \vdots & & & & \vdots \\ r_{m1}^{(N)} & r_{m2}^{(N)} & \cdots & r_{mm}^{(N)} & r_{m\,m+1}^{(N)} \\ r_{m+1\,1}^{(N)} & r_{m+1\,2}^{(N)} & \cdots & r_{m+1\,m}^{(N)} & r_{m+1\,m+1}^{(N)} \end{bmatrix} \tag{3-27}$$

因为在对 R 施行的 N 步变换中，元素 $r_{m+1\,m+1}$ 始终不是主元，所以每一步对 $r_{m+1\,m+1}$ 所用的变换公式就是普通的消元法；再由性质 2 和 3 知

$$r_{m+1\,m+1}^{(N)}=r'^{(l)}_{m+1\,m+1}=|\tilde{R}|/\tilde{R}_{m+1\,m+1} \tag{3-28}$$

我们有了(3-23)、(3-28)就可计算剩余平方和 $Q_1{}^{(N)}$ 和方差贡献 $V_{K_a}^{(N)}$ 了。

设逐步回归进行了 N 步，共引入 l 个变量 x'_{k_1}、$x'_{k_1}\cdots x'_{k_l}$。N 步的回归方程是

$$\hat{y}'=b_{k_1}{}'^{(N)}x'_{k_1}+b_{k_2}{}'^{(N)}x'_{k_2}+\cdots+b_{k_l}{}'^{(N)}x'_{k_l} \tag{3-29}$$

把方程组(3-22)看作方程组(3-18)的一个子方程组，则 N 步回归方程的回归系数 $b_{k_i}'^{(N)}$ 就是方程组(3-22)的解。

1. *回归方程式(3-29)的剩余平方和 $Q_1{}^{(N)}$*

$$\begin{aligned} Q_1{}^{(N)} &= r_{m+1\,m+1}-\sum_{i=1}^{l}b'_{k_i}r_{k_i m+1}=r_{m+1\,m+1}+\sum_{i=1}^{l}\frac{\tilde{R}_{m+1\,k_i}}{\tilde{R}_{m+1\,m+1}}r_{k_i m+1} \\ &= \left(r_{m+1\,m+1}\tilde{R}_{m+1\,m+1}+\sum_{i=1}^{l}\tilde{R}_{m+1\,k_i}r_{m+1\,k_i}\right)/\tilde{R}_{m+1\,m+1} \\ &= |\tilde{R}|/\tilde{R}_{m+1\,m+1}=r_{m+1\,m+1}^{(N)} \end{aligned} \tag{3-30}$$

2. *变量 x'_{k_a} 的方差贡献 $V_{k_a}{}^{(N)}$*

在第 N 步的基础上，第 $N+1$ 步拟引入 $x'_{k_a}(k_a\overline{\in}(k_1,k_2,\cdots,k_l)$，变量 x'_{k_a} 在此时的方差贡献可以看成从 $N+1$ 步回归方程中去掉变量 x'_{k_a}(得 N 步回归方程)剩余平方和的增加量，所以 x'_{k_a} 的方差贡献

$$\begin{aligned} V_{k_a}{}^{(N)} &= Q_1{}^{(N)}-Q_1{}^{(N+1)}=r_{m+1\,m+1}^{(N)}-r_{m+1\,m+1}^{(N+1)} \\ &= r_{m+1\,m+1}^{(N)}-(r_{m+1\,m+1}^{(N)}-r_{m+1\,k_a}^{(N)}\cdot r_{k_a m+1}^{(N)}/r_{k_a k_a}^{(N)}) \\ &= r_{m+1\,k_a}^{(N)}\cdot r_{k_a m+1}^{(N)}/r_{k_a k_a}^{(N)} \end{aligned} \tag{3-31}$$

在第 N 步的基础上，若第 $N+1$ 步拟剔除变量 $x'_{k_a}(k_a\in(k_1,k_2,\cdots,k_l)$，这时变量 x'_{k_a} 对 y 的方差贡献应是

$$Q_1{}^{(N+1)}-Q_1{}^{(N)}=r_{m+1\,m+1}^{(N+1)}-r_{m+1\,m+1}^{(N)}=-V_{k_a}^{(N)} \tag{3-32}$$

第 $N+1$ 步是引入变量时，x'_{k_a}，x'_{m+1} 都不在 N 步回归方程中，即 x'_{k_a}、$x'_{m+1}\overline{\in}(x'_{k_1},x'_{k_2},\cdots,x'_{k_l})$。由性质 1 知 $r_{m+1\,k_a}^{(N)}=r_{k_a m+1}^{(N)}$ 故 $V_{k_a}^{(N)}>0$。$N+1$ 步剔除时，x'_{k_a} 在 N 步回归方程中，x'_{m+1} 不在 N 步回归方程中，即 $x'_{k_a}\in(x'_{k_1},x'_{k_2},\cdots,x'_{k_l})$，而 $x'_{m+1}\overline{\in}(x'_{k_1},x'_{k_2},\cdots,x'_{k_l})$。由性质(1)知 $r_{m+1\,k_a}^{(N)}=-r_{k_a m+1}^{(N)}$ 故 $V_{k_a}^{(N)}<0$。为此我们约定：

引入变量 x'_{k_a} 时，其方差贡献为正；

剔除变量 x'_{k_a} 时，其方差贡献为负。

有了以上约定，在第 N 步基础上不论 $N+1$ 步是引入还是剔除变量 x'_{k_a} 的方差贡献就可写成一个式子，即

$$V_{k_a}^{(N)}=\frac{r_{m+1\,k_a}^{(N)}\cdot r_{k_a m+1}^{(N)}}{r_{k_a k_a}^{(N)}} \tag{3-33}$$

3. F 统计量

知道了 N 步回归方程的剩余平方和及 $N+1$ 步变量 x'_{k_a} 的方差贡献，就很容易写出 F 统计量。

根据式(3-16)，则

$$F_{k_a}=\frac{V_{k_a}^{(N)}(n-l-2)}{r_{m+1\,m+1}^{(N)}-V_{k_a}^{(N)}}\sim F(1,n-l-2)\text{分布}$$

根据式(3-17)，则

$$F'_{k_a}=\frac{|V_{k_a}^{(N)}|(n-l-1)}{r_{m+1\,m+1}^{(N)}}\sim F(1,n-l-1)\text{分布。}$$

其中 l 为第 N 步方程含有的自变量个数。

五、逐步回归的计算步骤

首先根据给定的数据

$$x_{1k},x_{2k},\cdots x_{mk},x_{m+1\,k} \qquad (k=1,2,\cdots,n)$$

按公式

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)(x_{jk}-\overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)^2\sum_{k=1}^{n}(x_{jk}-\overline{x}_j)^2}}$$

$$(i、j=1,2,3,\cdots,m+1)$$

计算相关系数增广矩阵 $R^{(0)}$。

第一步：选择第一个变量进入回归方程。

对 $i=1,2,\cdots,m$ 计算 x'_i 的方差贡献

$$V_i^{(0)}=r_{m+1\,i}^{(0)}r_{i\,m+1}^{(0)}/r_{ii}^{(0)}$$

选出其中最大的一个，记为 $V_{k_1}^{(0)}$，即

$$V_{k_1}^{(0)}=\max_{1\leqslant i\leqslant m}V_i^{(0)}$$

作 F 检验，有两种情况：

(1) 若

$$F_{k_1}=\frac{(n-2)V_{k_1}^{(0)}}{r_{m+1\,m+1}^{(0)}-V_{k_1}^{(0)}}>F^*$$

则变量 x'_{k_1} 对 y 的变化起着重要作用，引变量 x'_{k_1} 进回归方程，并按公式(3-21)变换 $R^{(0)}$ 得到第一步相关增广矩阵 $R^{(1)}$。

(2) 若

$$F_{k_1}=\frac{(n-2)V_{k_1}^{(0)}}{r_{m+1\,m+1}^{(0)}-V_{k_1}^{(0)}}\leqslant F^*$$

变量 x'_{k_1} 不能引入回归方程，回归就此结束，结果是回归方程中没引入任何变量。

第二步：对第一步中的情况(1)，需要继续下一步回归运算。

(1) 首先检验进入回归方程的变量是否有要剔除的。

由于回归方程中仅有一个变量 x'_{k_1}，并且是刚刚引入的，因此不可能立即被剔除，这步检验可省略。

(2) 检验是否要引入新变量

利用 $R^{(1)}$ 计算不在回归方程中变量的方差贡献 $V_i^{(1)}$：

$$V_i^{(1)}=r_{m+1\,i}^{(1)}r_{i\,m+1}^{(1)}/r_{ii}^{(1)},\quad 1\leqslant i\leqslant m\quad i\neq k_1$$

选出其中一个最大的，记为 $V_{k_2}^{(1)}$，即

$$V_{k_2}^{(1)}=\max_{\substack{1\leqslant i\leqslant m\\ i\neq k_1}}V_i^{(1)}$$

如果

$$F_{k_2}=\frac{(n-3)V_{k_2}^{(1)}}{r_{m+1\,m+1}^{(1)}-V_{k_1}^{(1)}}>F^*$$

把变量 x'_{k_2} 再引入回归方程，并按公式(3-21)变换 $R^{(1)}$ 得到第二步的相关增广矩阵 $R^{(2)}$。

(3) 如果

$$F_{k_2}=\frac{(n-3)V_{k_2}^{(1)}}{r_{m+1\,m+1}^{(1)}-V_{k_1}^{(1)}}\leqslant F^*$$

则逐步回归结束。由 $R^{(1)}$ 得到：

标准回归系数 $b'_{k_1}=r_{k_1\,m+1}^{(1)}$；

标准化变量回归方程 $\hat{y}'=b'_{k_1}x'_{k_1}=r_{k_1\,m+1}^{(1)}x'_{k_1}$；

剩余平方和 $Q_1^{(1)}=r_{m+1\,m+1}^{(1)}$，自由度 $f_{Q_1^{(1)}}=n-2$；

回归平方和 $Q_2^{(1)}=1-Q_1^{(1)}$，自由度 $f_{Q_2^{(1)}}=1$；

复相关系数 $R^{(1)}=(1-Q_1^{(1)})^{\frac{1}{2}}$。

第三步：对于第二步中的(2)，要作下一步回归运算：

(1) 首先检验已进入回归方程的变量是否有的要剔除

因为 x'_{k_2} 是上一步刚引入的，不可能马上又被剔除，因此只检验 x'_{k_1}。利用 $R^{(2)}$ 计算 x'_{k_1} 方差贡献。

$$V_{k_1}^{(2)}=r_{m+1\,k_1}^{(2)}r_{k_1\,m+1}^{(2)}/r_{k_1k_1}^{(2)}$$

如果

$$F'_{k_1}=\frac{(n-3)\,|V_{k_1}^{(2)}|}{r_{m+1\,m+1}^{(2)}}\leqslant F^*$$

则从回归方程中易除 x'_{k_1}，并对 $R^{(2)}$ 按式(3-21)进行变换，得 $R^{(3)}$。

(2) 如果

$$F'_{k_1}=\frac{(n-3)\,|V_{k_1}^{(2)}|}{r_{m+1\,m+1}^{(2)}}>F^*$$

则变量 x'_{k_1} 保留在回归方程中。此时再考虑是否还要引入变量，利用 $R^{(2)}$ 对不在回归方程中的变量计算方差贡献

$$V_i^{(2)}=r_{m+1\,i}^{(2)}r_{i\,m+1}^{(2)}/r_{ii}^{(2)},\quad 1\leqslant i\leqslant m,i\neq k_1,k_2$$

选出其中一个最大者，记为 $V_{k_3}^{(2)}$，即

$$V_{k_3}^{(2)}=\max_{\substack{1\leqslant i\leqslant m\\ i\neq k_1,k_2}} V_i^{(2)}$$

如果

$$F_{k_3}=\frac{(n-4)V_{k_3}^{(2)}}{r_{m+1\,m+1}^{(2)}-V_{k_3}^{(2)}}>F^*$$

就引人变量 x'_{k_3}，对 $R^{(2)}$ 按公式(3-21)进行一次变换，得到第三步的相关增广矩阵 $R^{(3)}$。

(3) 如果

$$F'_{k_1}=\frac{(n-3)V_{k_1}^{(2)}}{r_{m+1\,m+1}^{(2)}}>F^*$$

同时

$$F_{k_3}=\frac{(n-4)V_{k_3}^{(2)}}{r_{m+1\,m+1}^{(2)}-V_{k_3}^{(2)}}\leqslant F^*$$

那么回归方程中即没有可剔除的变量，又没有可引入回归方程的变量，逐步回归到此结束。由 $R^{(2)}$ 得以下结果：

标准回归系数 $b_{k_i}'^{(2)}=r_{k_i m+1}^{(2)}, i=1,2$；

标准化变量回归方程 $\hat{y}'=r_{k_1 m+1}^{(2)}x'_{k_1}+r_{k_2 m+1}^{(2)}x'_{k_2}$；

剩余平方和 $Q_1^{(2)}=r_{m+1\,m+1}^{(2)}, f_{Q1}^{(2)}=n-3$；

回归平方和 $Q_2^{(2)}=1-r_{m+1,m+1}^{(2)}, f_{Q_2}^{(2)}=2$；

复相关系数 $R^{(2)}=(1-r_{m+1\quad m+1}^{(2)})^{\frac{1}{2}}$。

⋮

⋮

假定回归已进了 N 步，引人 $x'_{k_1},x'_{k_2},\cdots,x'_{k_l}$ 共 l 个变量，回归还未结束，则需要在第 N 步的基础上进行第 $(N+1)$ 步回归运算。

第 $N+1$ 步运算：

(1) 首先检验进人回归方程的变量是否要剔除。对 $i=1,2,\cdots,l$，计算 x'_{k_i} 的方差贡献

$$V_{k_i}^{(N)}=r_{m+1\,k_i}^{(N)}r_{k_i m+1}^{(N)}/r_{k_i k_i}^{(N)}$$

设 $V_{k_a}^{(N)}$ 是 $V_{k_i}^{(N)}$ 中最小的一个，即

$$V_{k_a}^{(N)}=\min_{i=1,2,\cdots,l} V_{k_i}^{(N)}$$

如果

$$F'_{k_a}=\frac{(n-l-1)\,|V_{k_a}^{(2)}|}{r_{m+1\,m+1}^{(N)}}\leqslant F^*$$

从回归方程中剔除变量 x'_{k_a}，并按公式(3-21)对 $R^{(N)}$ 施行一次变换，得到 $R^{(N+1)}$。x'_{k_a} 剔除后，看是否还有要剔除的，若有重复(1)。

(2) 如果

$$F'_{k_a}=\frac{(n-l-1)\,|V_{k_a}^{(2)}|}{r_{m+1\,m+1}^{(N)}}>F^*$$

则回归方程中的 l 个变量无一可被剔除。此时考虑继续引人新变量。利用 $R^{(N)}$ 对不在回归方程中的变量计算方差贡献

$$V_i^{(N)}=r_{m+1\,i}^{(N)}r_{i\,m+1}^{(N)}/r_{ii}^{(N)} \qquad i\neq k_1,k_2,\cdots,k_l$$

取其最大者，记为 $V_{k_l+1}^{(N)}$，即

$$V_{k_l+1}^{(N)}=\max_{\substack{1\leqslant i\leqslant m\\ i\neq k_1,k_2,\cdots,k_l}} V_i^{(N)}$$

如果

$$F_{k_l+1}=\frac{(n-l-2)V_{k_l+1}^{(N)}}{r_{m+1\,m+1}^{(N)}-V_{k_l+1}^{(N)}}>F^*$$

则把变量 x'_{k_l+1} 引入回归方程，并对 $R^{(N)}$ 按公式(3-21)施行变换，得到 $R^{(N+1)}$。

(3) 如果

$$F_{k_l+1}=\frac{(n-l-2)V_{k_l+1}^{(N)}}{r_{m+1\,m+1}^{(N)}-V_{k_l+1}^{(N)}}\leqslant F^*，而$$

$$F'_{k_e}=\frac{(n-l-1)|V_{k_e}^{(N)}|}{r_{m+1\,m+1}^{(N)}}>F^*$$

则第 N 步回归方程中无变量可剔除，而 $N+1$ 步又无新变量引入，逐步回归停止。回归的主要结果由 $R^{(N)}$ 给出：

标准回归系数 $b_{k_i}^{\prime(N)}=r_{k_i m+1}^{(N)}, i=1,2,\cdots,l$；

回归方程 $\hat{y}'=r_{k_1 m+1}^{(N)}x'_{k_1}+r_{k_2 m+1}^{(N)}x_{k_2}+\cdots+r_{k_l m+1}^{(N)}x'_{k_l}$；

剩余平方和 $Q_1^{(N)}=r_{m+1\,m+1}^{(N)}, f_{Q1}^{(N)}=n-l-1$；

回归平方和 $Q_2^{(N)}=1-r_{m+1\,m+1}^{(N)}, f_{Q_2}^{(N)}=l$；

复相关系数 $R^{(N)}=(1-r_{m+1\,m+1}^{(N)})^{\frac{1}{2}}$。

对于第 $N+1$ 步中的(1)和(2)，要继续进行 $N+2$ 步回归运算。照上述方法做下去，直到回归方程中既没有对 y 作用不显著的变量要剔除，又没有对 y 作用显著的变量引入回归方程时，逐步回归最后结束。

在逐步回归中，用矩阵 $S=(S_{ij})(i,j=1,2,\cdots,m+1)$ 代替相关系数增广矩阵 $R=(r_{ij})(i,j=1,2,\cdots,m+1)$，逐步回归经过 N 步，共引入 l 个变量 $x_{k_1}x_{k_2}\cdots,x_{k_l}$，回归系数

$$b_{k_i}^{(N)}=\frac{\sigma_{k_i}\sigma_{m+1}}{\sigma_{k_i}\sigma_{k_i}}r_{k_i m+1}^{(N)}=\frac{\sigma_{m+1}}{\sigma_{k_i}}r_{k_i m+1}^{(N)}$$

$$b_0=\bar{y}-\sum_{i=1}^{l}b_{k_i}^{(N)}\bar{x}_{k_i}$$

于是得原变量的回归方程

$$\hat{y}=b_0+b_{k_1}^{(N)}x_{k_1}+b_{k_2}^{(N)}x_{k_2}+\cdots+b_{k_l}^{(N)}x_{k_l}$$

§4 逐步回归 FORTRAN 源程序

在给定的 F 统计检验量下，该程序逐步选人或剔除变量建立回归方程式并对给定样品进行验算。现将程序中的主要参数、符号及程序的使用方法说明如下：

一、符号说明

x(i,j)——i 行 j 列的二组数组。其中 i 为样品号下标，j 为变量号下标(最后一个变量为因变量 y)；

rs——存放相关系数的数组；

ax——存放变量平均值的数组名；

segm——存放变量标准方差数组名；

f1 与 f2——给定的引入剔除 F 检验值；

f11，f22——各为计算的引入和剔除变量的 F 检验统计量；

ll、kk——各为引入与未引入回归方程的变量号的存储单元，按引入与剔除顺序存入；

n——样品数，即观测值组数；

m——自变量数(第 m+1 个变量为因变量 y)；

b——存放回归系数的数组名；

b0——回归方程的常数项；

r——复相关系数；

yy——计算回归系数及样品回归值子程序；

pp——变量平均值及标准方差子程序；

rcoef——计算相关系数矩阵子程序；

im——计算方差贡献子程序；

ma——矩阵变换子程序。

二、程序使用说明

1. 数据文件

在使用本程序前，需先把样品观测值建立一个 n 行 m+1 列的数据文件。变量在文件中的存放顺序是从左至右依次为 $x_1, x_2, \cdots, x_m, x_{m+1}$($x_{m+1}$是因变量 y)。

2. 程序运行

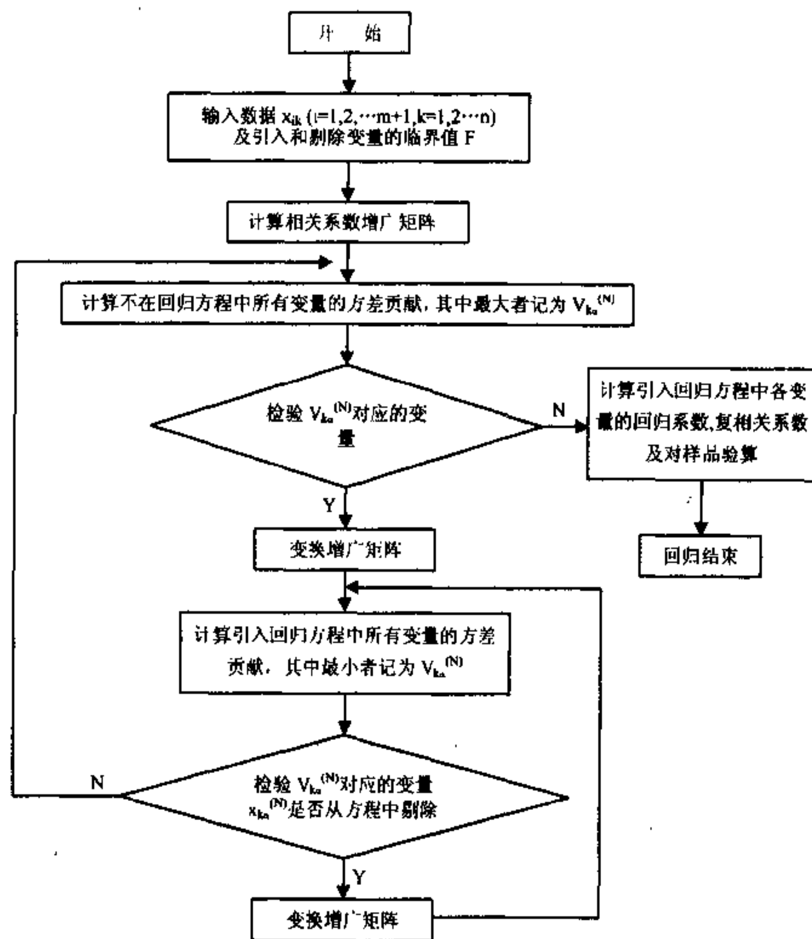
程序运行时由键盘输入样品数 n、变量数 m，给定的引入 F 检验值 f1 和剔除的 F 检验值 f2 及存放变量观测值和计算结果的数据文件名。

3. 主要输出结果

输出的主要结果是引入回归方程中各变量的回归系数、回归方程的常数项及复相关系数等。

三、源程序

图 3-3 是逐步回归分析流程。



开 始
输入数据 x_{ik} (i=1,2,…m+1,k=1,2…n)
及引入和删除变量的临界值 F
计算相关系数增广矩阵
计算不在回归方程中所有变量的方差贡献，其中最大者记为 $V_{ka}^{(N)}$
检验 $V_{ka}^{(N)}$对应的变量
N
计算引入回归方程中各变量的回归系数,复相关系数及对样品验算
回归结束
Y
变换增广矩阵
计算引入回归方程中所有变量的方差贡献，其中最小者记为 $V_{ka}^{(N)}$
N
检验 $V_{ka}^{(N)}$对应的变量 $x_{ka}^{(N)}$是否从方程中剔除
Y
变换增广矩阵

图 3-3 逐步回归流程图

```
c          program—zbhg.for
$debug
           dimension x(18,3),ax(3),segm(3),rs(3,3)
           dimension v(3),b(3),y(18),ll(3),kk(3),sem(18)
           character fam * 10
           character jieguo * 10
           write( * , * ) 'please input n,m,f1,f2'
           read ( * , * ) n,m,f1,f2
           write( * , * ) 'Enter your dat and jieguo filename '
           read ( * ,'(a)') fam,jieguo
           m1=m
           m=m+1
           open (3,file=fam) ! ' ',status='old')
           open (2,file=jieguo)
           do 10 i=1,n
           read(3, * ) (x(i,j),j=1,m)
10         continue
           call pp(x,ax,segm,n,m)
           call rcoef(x,rs,ax,segm,n,m,m1)
           do 12 k=1,m1
           li(k)=0
12         kk(k)=k
           m2=0
15         vmax=0.
           m3=m1-m2
           call im(rs,kk,v,m3,m,m1,vmax,l1)
           write(2,20) l1,vmax
20         format(4x,'l1=',i2,4x,'vmax=',f10.4)
           f11=vmax * float(n-m2-2)/(rs(m,m)-vmax)
           if(f11.le.f1) go to 80
           write(2,21) f11,l1,vmax
21         format(//,26x,'f11=',f10.4,4x,'l1=',i3,4x,'vmax=',e10.4)
           ll(m2+1)=l1
           do 25 i=1,m3
           ii=i
           if(kk(i).eq.l1) go to 30
25         continue
30         if(ii.eq.m3+1) go to 35
           kk(ii)=kk(ii+1)
           ii=ii+1
```

```
          go to 30
35        m2=m2+1
          l=l1
          call ma(rs,l,m)
          if(m2.eq.m1) go to 80
40        vmin=-1.e15
          call im(rs,ll,v,m2,m,m1,vmin,l2)
          write(2,45) l2,vmin
45        format(4x,'l2=',i2,4x,'vmin=',f10.4)
          f21=-vmin*float(n-m1-1)/rs(m,m)
          if(f21.le.f2) go to 50
          go to 15
50        write(2,55) f21,l2,vmin
55        format(1h0,26x,'f21=',f10.4,4x,'l2=',i3,4x,'vmin=',e10.4)
          do 60 i=1,m2
          if(ll(i).eq.l2) go to 65
60        continue
65        m4=i
70        if(m4.eq.m2+1) go to 75
          ll(m4)=ll(m4+1)
          m4=m4+1
          go to 65
75        m2=m2-1
          kk(m1-m2)=l2
          l=l2
          call ma(rs,l,m)
          if(m2.eq.m1) go to 80
          go to 40
80        call yy(rs,x,ax,segm,sem,ll,y,b,m,m1,m2,n)
          write(2,85)
85        format(//,7x,'no',12x,'y(j)',11x,'y0(j)',10x,'sem(j)')
          write(*,*)
          do 95 i=1,n
          write(2,90) i,y(i),x(i,m),sem(i)
90        format(4x,i5,3(6x,f10.4))
95        continue
          stop
          end

          subroutine yy(rs,x,ax,segm,sem,ll,y,b,m,m1,m2,n)
```

```
      dimension rs(m,m),x(n,m),ax(m),segm(m),sem(n)
      dimension ll(m1),y(n),b(m1)
      ss=0.
      do 20 i=1,m2
      l=ll(i)
      b(l)=rs(l,m) * segm(m)/segm(l)
      write(2,10) l,b(l)
10    format(10x,'b',i2,'=',e12.4)
20    ss=ss+b(l) * ax(l)
      b0=ax(m)-ss
      r=sqrt(1-rs(m,m))
      write(2,30) b0,r
30    format(10x,'b0=',e13.4/10x,'r= ',e13.4)
      do 50 j=1,n
      y(j)=0.
      do 40 i=1,m2
      l=ll(i)
40    y(j)=y(j)+b(l) * x(j,l)
      y(j)=y(j)+b0
50    continue
      do 60 j=1,n
      sem(j)=x(j,m)-y(j)
60    continue
      return
      end

      subroutine pp(x,ax,segm,n,m)
      dimension x(n,m),ax(m),segm(m)
      an=n
      do 20 i=1,m
      s1=0.
      s2=0.
      do 10 j=1,n
      s1=s1+x(j,i)
      s2=s2+x(j,i) * * 2
10    continue
      ax(i)=s1/an
      s=s2-an * ax(i) * * 2
      segm(i)=sqrt(s)
20    continue
```

```
      return
      end

      subroutine rcoef(x,rs,ax,segm,n,m,m1)
      dimension x(n,m),rs(m,m),ax(m),segm(m)
      do 10 i=1,m1
      j1=i+1
      do 10 j=j1,m
      s=0.
      do 20 k=1,n
      s=s+(x(k,i)-ax(i)) * (x(k,j)-ax(j))
20    continue
      rs(i,j)=s/(segm(i) * segm(j))
      rs(j,i)=rs(i,j)
10    continue
      do 30 i=1,m
      rs(i,i)=1.0
30    continue
      do 40 i=1,m
40    write(2,50) (rs(i,j),j=1,m)
50    format(1x,10f8.4)
      return
      end

      subroutine im(rs,kk,v,mi,m,m1,vmai,lk)
      dimension rs(m,m),kk(m1),v(m1)
      do 40 i=1,mi
      k=kk(i)
      v(k)=rs(k,m) * rs(m,k)/rs(k,k)
      write(2,50) v(k)
50    format(30x,'v=',e12.4)
      if(v(k).le.vmai) go to 40
      vmai=v(k)
      lk=k
40    continue
      return
      end

      subroutine ma(rs,l,m)
      dimension rs(m,m)
```

```
      if(rs(l,l).lt.1.e-5) write(2,10) l
10    format(1x,'l=',i6)
      c1=1.0/rs(l,l)
      do 30 i=1,m
      if(i.eq.l) go to 30
      c2=rs(i,l)*c1
      do 20 j=1,m
      if(j.ne.l) rs(i,j)=rs(i,j)-rs(l,j)*c2
20    continue
30    continue
      do 40 j=1,m
      if(j.eq.l) go to 40
      rs(l,j)=rs(l,j)*c1
      rs(j,l)=-rs(j,l)*c1
40    continue
      rs(l,l)=c1
      r=sqrt(1.0-rs(m,m))
      return
      end
```

§5 应用算例

【例 1】 镜质体反射率 R_O 是衡量生油岩成熟度的一个重要指标。据胜利油田地质研究院资料(表 3-1)计算，东营凹陷、沾化凹陷沙河街组 R_O 实测平均值 $\overline{R}_O$ 与埋藏深度 H 具有良好的线性关系，如图 3-4 所示。以下两式

$$\overline{R}_O = 0.0002H - 0.0115 \qquad \overline{R}_O = 0.0002H + 0.09521$$

分别是东营凹陷和沾化凹陷沙河街组 H 对 $\overline{R}_O$ 的回归方程，其相关系数均等于 0.95。这一关系表明，可利用生油层的埋藏深度寻找研究区内的有机成熟区。又如，松辽盆地南部 61 个样品的镜质体反射率 R_O 与时间-温度指数 TTI 之间也有密切的相关关系，如图 3-5 所示。其相关系数为 0.99 的一元回归方程为：

$$R_O = 0.493\mathrm{Lg}TTI + 0.219$$

表 3-1 镜质体反射率与埋深数据

埋深 /m	R_O 实测平均值/%、回归值/%及偏差					
	东营凹陷			沾化凹陷		
	实测平均值	回归值	偏差	实测平均值	回归值	偏差
1200	0.29	0.22	0.07	0.30	0.27	0.03
1300	0.30	0.24	0.06	0.32	0.28	0.04
1400	0.31	0.26	0.05	0.33	0.30	0.03
1500	0.32	0.28	0.04	0.33	0.32	0.01
1600	0.33	0.30	0.03	0.34	0.33	0.01
1700	0.34	0.32	0.02	0.35	0.35	0.00
1800	0.35	0.34	0.01	0.37	0.37	0.00
1900	0.36	0.36	0.00	0.38	0.38	0.00
2000	0.38	0.38	0.00	0.41	0.40	0.01
2100	0.39	0.39	0.00	0.44	0.42	0.02
2200	0.40	0.41	0.01	0.46	0.44	0.02
2300	0.41	0.43	−0.02	0.47	0.45	0.02
2400	0.42	0.45	−0.03	0.48	0.47	0.01
2500	0.43	0.47	0.04	0.49	0.49	0.00
2600	0.43	0.49	−0.06	0.49	0.50	0.01
2700	0.44	0.51	−0.07	0.49	0.52	0.03
2800	0.46	0.53	−0.07	0.50	0.54	−0.04
2900	0.47	0.55	−0.08	0.50	0.55	0.05
3000	0.49	0.57	−0.08	0.50	0.57	0.07
3100	0.52	0.59	−0.07	0.51	0.59	0.08
3200	0.55	0.61	−0.06	0.52	0.60	0.08
3300	0.58	0.63	−0.05	0.55	0.62	0.07
3400	0.61	0.65	−0.04	0.58	0.61	−0.06
3500	0.65	0.67	−0.01	0.62	0.65	−0.03
3600	0.68	0.68	0.00	0.67	0.67	0.00
3700	0.72	0.70	0.02	0.70	0.69	0.01
3800	0.77	0.72	0.05	0.75	0.71	0.01
3900	0.82	0.74	0.08	0.79	0.72	0.07
4000	0.88	0.76	0.12	0.83	0.74	0.09
4100	0.92	0.78	0.14	0.87	0.76	0.11

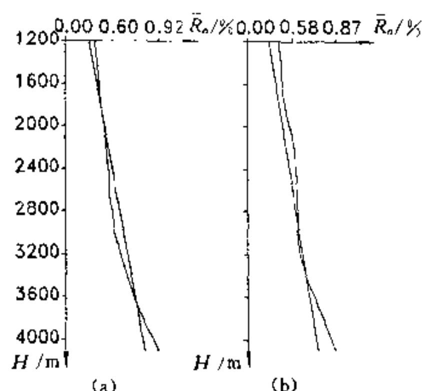


图 3-4 $\bar{R}_o$ 与 H 关系图
(a) 东营凹陷，(b) 沾化凹陷

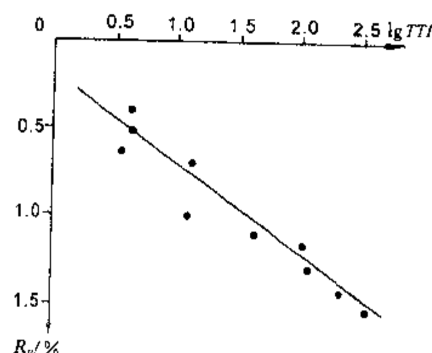


图 3-5 R_o 与 $LgTTI$ 之间的线性关系图

【例 2】 N·N 涅斯乔夫等根据世界上勘探程度较高的二十二个含油气盆地的资料，认为盆地的油气地质储量与盆地的平均体积速度有关，并求得

$$\lg Q = 1.613V + 2.183$$

的回归方程。

式中 Q——油气地质储量，Mt；

V——盆地沉积物的平均体积速度，$10^3 km^3/Ma$。

另外，根据统计分析，上述盆地中大油气田（地质储量大于 2 亿吨，或者地质储量虽不足 2 亿吨，但占全区总储量 10%以上的油气田）的累积储量 q 与全区油气总储量的关系为：

$$\lg q = 1.196\lg Q - 1.149$$

因此，用体积速度方程不仅可以预测评价区的总储量，而且还可以估计该区大油气田的总储量。

【例 3】 我国东部地区一些勘探程度较高的含油气凹陷的单位面积内的油气储量（或资源量）Q 与 Vf、H、Vr、Sn 及 N 有着密切关系。

1985 年 2 月，朱子仁等采用探明储量建立回归方程：

$$Q = 0.136Vf + 0.729H + 0.356Vr + 0.152Sn - 0.12N - 5.37$$

式中 Q——单位面积内的油气储量，$10^4 t/km^2$；

Vf——生油岩体积与沉积岩体积之比，%；

H——总烃与有机碳之比（有机质转化率），%；

Vr——储集岩体积与沉积岩体积之比，%；

Sn——近油源圈闭面积与沉积岩面积之比，%；

N——含油气凹陷经历的剥蚀次数。

采用概算储量得到的回归方程为：

$$Q = 0.835Vf + 0.597H + 0.269Vr + 0.142Sn - 0.5N - 6.654$$

根据概算储量回归方程，预测前梨园洼陷沙二下段 460 平方公里内的石油资源总量为：

$$\sum Q = 28.834 \times 460 = 13263.64(10^4 t)。$$

【例 4】 康南(Connan)用表 3-2 中前 12 个地区的生油层数据回归出大量生油门限时间

方程式

$$\ln t=6942\frac{1}{T+273}-14.965 \tag{3-34}$$

在前十二个地区生油层数据的基础上，增加我国六组生油层数据（表 3-2 中后六行数据），取实际深度 H、现在温度 T 及 $1/(T+273)$、$1/H$、H^2 和 T^2 共六个地质变量，以地层年龄 t 作为因变量，进行多元逐步回归分析。结果选入 $1/(T+273)$ 及 $1/H$ 两个变量，得出回归方程式：

$$\ln t=7585\frac{1}{T+273}-2370\frac{1}{H}-15.95 \tag{3-35}$$

复相关系数 $r=0.94$。

表 3-2　十八个盆地（地区）的生油层数据

序号	盆地（地区）	门限值		
		地层年龄(t)/Ma	现在温度(T)/℃	实际深度(H)/m
1	杜阿拉盆地（喀麦隆）	70	65	1200
2	落山矶盆地（美国）	12	115	2440
3	文吐拉盆地（美国）	12	127	2740
4	巴黎盆地（法国）	180	60	1400
5	阿启坦盆地(1)（法国）	112	90	3300
6	阿启坦盆地(2)（法国）	135	72	2500
7	卡马尔圭盆地（法国）	38	106	3250
8	阿尤恩地区	105	85	2740
9	苏禄海盆地（沙巴）	12	120	3050
10	塔拉纳基盆地（新西兰海上）	70	80	2900
11	亚马逊盆地（委内瑞拉）	359	62	1750
12	塔拉纳基盆地（新西兰海上）	32	95	3350
13	东营盆地	35	93	2200
14	潜江盆地	35	90	2200
15	松辽盆地(1)	110	70	1330
16	松辽盆地(2)	100	65	1230
17	松辽盆地(3)	90	63	1180
18	辽河盆地	50	81	1700

分析表 3-2 中的数据，舍去特殊的 5 号和 11 号两个样品，得回归方程式：

$$\ln t=6708\frac{1}{T+273}-1717\frac{1}{H}-13.91 \tag{3-36}$$

复相关系数 $r=0.96$。

把表 3-2 中的地层温度和实际深度分别代入式(3-34)、(3-35)及式(3-36)，计算大量生油门限时间，并按

$$\left|\frac{\text{生油门限时间}-\text{实际地层年龄}}{\text{实际地层年龄}}\right|\times 100\%$$

计算误差，结果列于表 3-3。

表 3-3 中的误差及平均误差表明，深度是研究有机质热深化成油时间不可忽视的因素。由式(3-35)或式(3-36)可以看出，生油层温度高，埋藏浅将缩短生油门限时间，反之生油门限时间长，这说明地温梯度也是影响生油门限时间的一个重要因素。另外，式(3-35)和式(3-36)的计算结果表明，异常数据对估计值有较大的影响。

表 3-3　生油门限时间回归值及误差

序　号	地层年龄/Ma	式(3-34)		式(3-35)		式(3-36)	
		回归值/Ma	误差/%	回归值/Ma	误差/%	回归值/Ma	误差/%
1	70	263.34	276	91.43	31	90.60	30
2	12	18.67	56	13.84	15	14.50	21
3	12	10.91	9	8.56	29	9.36	22
4	180	358.47	99	169.81	6	150.07	17
5	112	64.01	43	68.51	39	59.74	47
6	135	173.60	29	161.95	20	127.73	5
7	38	28.55	25	28.05	26	26.16	31
8	105	83.60	20	79.20	25	66.95	36
9	12	14.87	24	13.11	9	13.45	12
10	70	110.03	57	112.15	60	90.38	29
11	359	316.52	10	207.93	42	441.21	23
12	32	49.36	54	52.13	63	45.10	41
13	35	54.72	56	40.31	15	38.12	9
14	35	64.01	83	47.84	37	44.35	27
15	110	195.21	77	79.96	27	78.19	29
16	100	263.34	163	95.95	4	94.03	6
17	90	297.59	231	101.06	12	99.74	11
18	50	104.08	108	59.28	19	56.40	13
平均误差/%			79		27		23

石油勘探实践证明，许多油气藏是生油岩中的油气就近运移而形成的，因此生油门限时间回归方程给我们提供了一个追踪生油面积的找油途径。也就是说，绘制大量生油门限时间与地层实际年龄的偏差等值线图，正偏差区即为石油热演化成熟区。

【例 5】 利用回归方程预测含油面积系数。陈立平、陈子思等，曾利用回归方程对四川盆地侏罗系自流井群大安寨组的含油面积系数进行预测，并作出了预测图，为油气资源评介提供了重要参数。其作法是将评价区分成大小为 10km×10km 的 675 个单元，取反映沉积、生油、构造三方面的 14 项地质参数为自变量，它们是：

构造因素：
- x_1—大安寨组底面构造七次趋势剩余值，m；
- x_2—早第三纪前大安寨组底面古构造六次趋势剩余值，m；
- x_3—大安寨组底面现今构造海拔高度，m；
- x_4—早第三纪前大安寨组底面古构造六次趋势值，m。

沉积因素：
- x_5—介屑灰岩、页岩沉积韵律数；
- x_6—页岩厚度/介屑灰岩厚度；
- x_7—(页岩厚度+介屑灰岩厚度)/组厚度%；
- x_8—页岩厚度/组厚×介屑灰岩厚度/组厚(小数)；
- x_{12}—页岩厚度，m；
- x_{13}—介屑灰岩厚度，m；
- x_{14}—介屑灰岩单层平均厚度，m；

生油因素：
- x_9—有机质成熟度时间温度指数(TTI)；
- x_{10}—单位面积生油量，t/km^2；
- x_{11}—单位体积生油量，t/km^3。

因变量 y——含油面积系数(单元中累计产油超过一吨的油井与总井数之比值(%)。由于多年的勘探开发及综合研究,这 15 项参数在每个单元上的取值,均可由绘制的相应等值线中求得。在 675 个单元中,有 139 个有钻探资料,加上边界控制点 11 个共 150 个单元。根据 150 个单元的数据,用逐步回归分析法,取 $f_2=f_2=2.5$,求得含油面积系数回归方程如下:

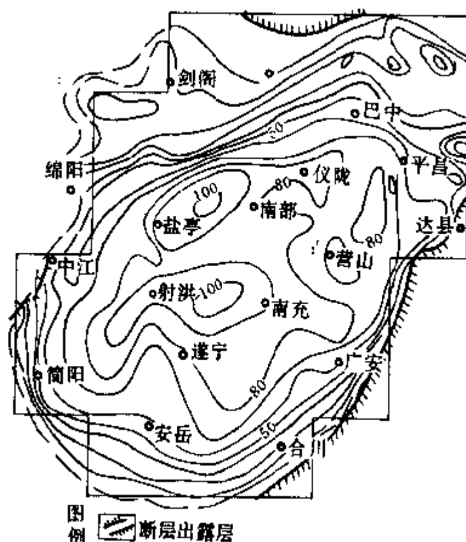


图 3-6 四川盆地侏罗系自流井群
大安寨组含油面积系数预测图
(回归方程标准方差为 14.35; 图中数值为百分数)

y(含油面积系数)$=0.780515+2.64\times10^{-4}x_1+6.82\times10^{-4}x_2-2.08\times10^{-4}x_3-3.4\times10^{-4}x_4+7.143\times10^{-3}x_5+2.8607\times10^{-2}x_6+8.442\times10^{-3}x_7$ 它的复相关系数 $R=0.8432$,方程高度显著。

在进入方程的自变量中,x_1、x_2、x_3、x_4 属于构造因素,x_5、x_6、x_7 属于沉积因素。它们的组合正反映了大安寨组是受构造、岩性岩相控制的裂缝性油藏的特点。

用回归方程求出每个评价单元的预测含油面积系数,最后,在此基础上作出整个评价区上的含油面积系数预测图,如图 3-6 所示。

习 题

1. 什么是回归分析?
2. 回归分析的研究对象是什么?
3. 在多元回归中,为何要对变量进行标准化?
4. 由标准化变量求得的回归方程具有什么样的形式?它的回归系数与原始变量回归方程的系数之间有何关系?
5. 什么是逐步回归分析?在逐步回归分析中如何对变量进行筛选?
6. 试述逐步回归分析的计算步骤。
7. 对例 4 中的数据,如果取检验水平 $a=0.1$,试用逐步回归程序求回归方程并对其进行显著性检验。

第四章　聚类分析

§1　聚类分析及聚类统计量

一、聚类分析

在地质学研究领域内，有很多分类或属于分类的问题。例如：沉积岩、古生物、矿物、油气地表化探指标、油气藏的分类等，都是直接分类的例子；含油气盆地的油气远景评价、油源对比等，是属于用分类方法解决的问题；地层的划分，也可以说是与上不同的另一种分类。

为了叙述方便，把分类问题中的行为对象统称为客体。在传统的分类中，往往是综合客体的大量资料，进行定性的分类。当考虑的因素较多时，往往会片面地强调某些因素，而忽视其它因素，同时又会因人的认识不同，使分类结果带有一定的片面性和主观臆断。

聚类分析又称点群分析。它是按照客体在性质上或成因上的亲疏关系，对客体进行定量分类的一种多元统计分析方法。这种分类方法不仅综合考虑了所有的因素，而且又不受已有分类结构的影响，只是以某种分类统计量为分类依据，对客体进行分类，因此这就有可能突破传统地质学建立的一些定性分类系统，而得到更合理的分类结果。

按照客体之间的关系，可把分类中的客体分为无序客体和有序客体。彼此之间没有次序约束关系的客体称为无序客体，反之，为有序客体。例如：对油气藏分类时，参与分类的油气藏就是无序客体；沿地层剖面按由老到新的顺序取了 n 个岩样，如果把岩样的分类结果用于地层划分，那么分类时，岩样的顺序是不能打乱的，这些岩样就是有序客体。对无序客体和有序客体的聚类分析又分别称为无序客体和有序客体聚类分析。

按照聚类分析方法原理，又可分为聚合法聚类分析和分解法聚类分析等。

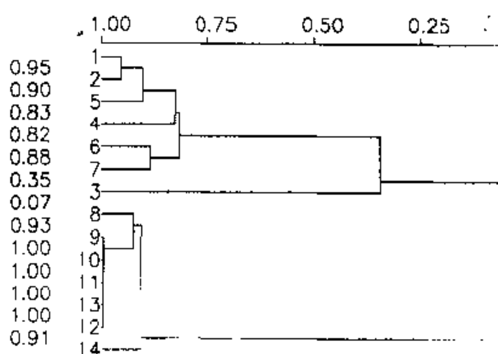


图 4-1　油气化探指标聚类谱系图
（图中 1，2，…，14 为化探指编号）

聚合法聚类分析，在开始时每个客体自成一类，如图 4-1 所示，然后以某种表示客体亲疏关系的分类统计量为分类依据，把一些彼此之间关系最亲密的客体聚集合并为一类，把另一些彼此之间亲近的客体聚合为另一类，…。在客体聚合为类（有的类内可能只有一个客体）的基础上，再根据类之间的亲疏程度继续合并，直到全部客体聚为一类为止，给出一个反映客体间亲疏关系的定量分类系统——聚类分析谱系图。以相关系数为分类统计量，对 14 项油气地表化探指标进行聚合法聚类分析的谱系图如图 4-1 所示。

在图 4-1 上，14 项指标明显地分为两大类，它们之间的相关系数为 0.07，反映了指标间的成因差别。

分解法聚类分析的分类过程与聚合法恰好相反，开始把全部客体看成一类，然后根据某种

统计准则进行分解(分类),一直分解到所需的分类为止。

常用聚合法聚类分析对无序客体分类,而分解法聚类分析则多用于对有序客体的分类。按照客体,聚类分析又分为 Q 型和 R 型聚类分析。前者对样品进行分类,而后者对变量分类。

二、聚类统计量

如果有 n 个样品,且每个样品包含 m 个变量,那么 n 个样品 m 个变量的观测值 $x_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,m)$ 构成一个 $n\times m$ 的数据矩阵 $X_{n\times m}$。

$$X_{n\times m}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

从 $X_{n\times m}$ 中可以看出:

① 矩阵的第 i 行表示第 i 个样品 m 个变量的观测值,可把第 i 行视为 m 维空间的一个点或一个矢量。

② 矩阵的第 j 列表示第 j 个变量的 n 次观测值,可把第 j 列视为 n 维空间中的一个点或一个矢量。

③ 由①、②可知,研究样品的相似性,把相似程度高的样品归为一类,就等价于研究矩阵行与行之间的关系,对矩阵的行进行归类。研究变量间的相关性,把相关程度高的变量归为一类,就等价于研究矩阵列与列之间的关系,对矩阵的列进行归类。

聚类统计量是指用于衡量客体间相似(或相关)程度的某种指标。

(一) Q 型聚类统计量

1. 相似系数

矢量 $X_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ 与矢量 $X_j=\{x_{j1},x_{j2},\cdots,x_{jm}\}$ 夹角的余弦定义为矢量 X_i 与 X_j 的相似系数,即

$$r_{ij}=\cos\theta_{ij}=\frac{X_iX_j}{|X_i||X_j|}=\frac{\sum_{k=1}^{m}x_{ik}\cdot x_{jk}}{\sqrt{\sum_{k=1}^{m}x_{ik}^2\sum_{k=1}^{m}x_{jk}^2}} \qquad (4\text{-}1)$$

$$(i,j=1,2,\cdots,n)$$

$[r_{ij}]_{n\times n}$ 是实对称矩阵,且 $r_{11}=r_{22}=\cdots r_{nn}=1$。$r_{ij}$ 愈接近于 1,样品 X_i 与 X_j 的性质愈相近,视 r_{ij} 的相对大小对样品分类。

2. 相关系数

$$R_{ij}=\frac{\sum_{k=1}^{m}(x_{jk}-\bar{x}_j)(x_{ik}-\bar{x}_i)}{\sqrt{\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)^2\sum_{k=1}^{m}(x_{jk}-\bar{x}_j)^2}} \qquad (4\text{-}2)$$

$$(i,j=1,2,\cdots,n)$$

相关系数 R_{ij} 反映 X_i 和 X_j 的相关程度,显然 R_{ij} 在区间[-1,1]内取值,且 $R_{ij}=R_{ji}$,$R_{ii}=1$。应把 R_{ij} 相对大的样品分为一类。

3. 距离系数

(1) 欧氏距离

在正交坐标系中，两个样品点 X_i 和 X_j 之间的距离为：

$$d_{ij}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2} \tag{4-3}$$

d_{ij}越小，样品 X_i 与 X_j 的性质愈相近。所以，应把 d_{ij}相对小的样品分为一类。为了不致使 d_{ij}太大(因为只是将 n 个样品中 d_{ij}相对小的样品分为一类，把 d_{ij}缩小同样倍数，不影响分类结果)，故将式(4-3)改写为：

$$d_{ij}=\sqrt{\frac{1}{m}\sum_{k=1}^{m}(x_{ik}-x_{jk})^2} \tag{4-4}$$

$$(i,j=1,2,\cdots,n)$$

$[d_{ij}]_{n\times n}$为实对称矩阵，且 $d_{11}=d_{22}\cdots=d_{nn}=0$。

(2) 斜交距离

在斜交坐标系中，两个样品点 X_i 和 X_j 之间的距离为：

$$Od_{ij}=\sqrt{\sum_{k=1}^{m}\sum_{l=1}^{m}(x_{ik}-x_{jk})(x_{il}-x_{jl})r_{kl}} \tag{4-5}$$

$$(i,j=1,2,\cdots,n)$$

式中

$$r_{kl}=\frac{\sum_{i=1}^{n}(x_{ik}-\overline{x}_k)(x_{il}-\overline{x}_l)}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\overline{x}_k)^2\sum_{i=1}^{n}(x_{il}-\overline{x}_l)^2}} \tag{4-6}$$

$$(k,l=1,2,\cdots,m)$$

(二) R 型聚类统计量

如前所述，研究变量间的亲缘关系，就是研究原始数据矩阵 $X_{n\times m}$列之间的关系，故只须对 Q 型聚类统计量略加修改就得 R 型聚类统计量。

1. 相似系数

$$r_{ij}=\frac{\sum_{k=1}^{n}x_{ki}\cdot x_{kj}}{\sqrt{\sum_{k=1}^{n}x_{ki}^2\cdot\sum_{k=1}^{n}x_{kj}^2}} \tag{4-7}$$

$$(i,j=1,2,\cdots,m)$$

2. 相关系数

$$R_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)^2\sum_{k=1}^{m}(x_{kj}-\overline{x}_j)^2}} \tag{4-8}$$

$$(i,j=1,2,\cdots,m)$$

3. 距离系数

(1) 欧氏距离

$$d_{ij}=\sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_{ki}-x_{kj})^2} \tag{4-9}$$

$$(i,j=1,2,\cdots,m)$$

(2) 斜交距离

$$Od_{ij}=\sqrt{\sum_{k=1}^{n}\sum_{l=1}^{n}(x_{ki}-x_{kj})(x_{li}-x_{lj})r_{kl}} \tag{4-10}$$

$$(i,j=1,2,\cdots,m)$$

式中

$$r_{kl}=\frac{\sum_{i=1}^{m}(x_{ki}-\bar{x}_k)(x_{li}-\bar{x}_l)}{\sqrt{\sum_{i=1}^{m}(x_{ki}-\bar{x}_k)^2\sum_{i=1}^{m}(x_{li}-\bar{x}_l)^2}} \tag{4-11}$$

$$(k,l=1,2,\cdots,n)$$

R 型聚类统计量是衡量变量相关程度的统计指标，根据它的相对大小对变量进行分类，进而研究它们之间的共生组合关系。

§2 聚合法聚类分析

聚合法是将客体类由多变少，直到把全部客体合并成一类的一种聚类分析方法。它是目前最常用的聚类分析方法。其聚类过程是：

① 开始聚类时，每个客体(样品或变量)自成一类；

② 按某种聚类统计量，计算客体间的亲疏关系，把最亲近的两个客体合并成一类，形成一个由两个客体为一类的客体集团(类)；

③ 计算某一类与其余各类之间的亲疏关系，把最亲近的两个类合并成新的一类；

④ 如果分类的总数目仍大于1，则对上一步中合并成的新类计算各类间的亲疏关系，按新类的亲近程度继续合并新类，直到把全部客体聚合为一个大类为止。

一、距离类统计量聚合法

对于距离类统计量来说，两个样品(或变量)点之间的距离，在概念上是明确的。但是，由若干样品(或变量)结合为一个样品(或变量)点集团，即一个类后，类与类之间的距离可有不同的定义方法。例如，可以定义类与类之间的距离为两类中相距最近的两个样品(或变量)点之间的距离；也可以把两类中相距最远的两个样品(或变量)点之间的距离作为两类间的距离；还可取两类的重心距离作为两类之间的距离等。由于类间距离的定义不同，也就产生了不同的聚合法。在此，介绍四种距离类度量聚合法。

(一) 最短距离法

若类 p 与 q 合并为 1 类，那么定义

$$D_{pq}=\min_{\substack{X_i\in q\\X_j\in p}} d_{ij} \tag{4-12}$$

为类 p 与 q 之间的最短距离。其中 d_{ij}是类 p 和类 q 中两个样品(或变量)点之间的距离。用距

离 D_{pq}进行的聚合聚类，叫做最短距离法。其聚类过程可归结为如下 4 步：

1. 准备工作

计算所有样品(或变量)点之间的距离，记为

$$D^{(0)}=[d_{ij}^{(0)}]_{r\times n}$$

$$(i,j=1,2,\cdots,n \text{ 或 } i,j=1,2,\cdots,m)$$

2. 第一级聚类

在矩阵 $D^{(0)}$中寻找非对角线上值为最小的对称元素，若为 d_{pq}，则将类 p 与 q 合并成一个新类 r，$r=\{p,q\}$。同时计算新类 r 与其他各类之间的距离：

$$D_{rk}=\min_{\substack{X_i\in r\\X_j\in k}} d_{ij} \tag{4-13}$$

$$(k=1,2,\cdots,n(\text{或 } m),k\neq r)$$

在这一级聚类中，不妨设 $p<q$，保留小号 p 作为新类 r 的类号。删去 $D^{(0)}$中第 p、q 行及第 p、q 列上的元素，并将刚计算出的距离 D_{rk}写在第 p 行 p 列上，所形成的距离矩阵记为 $D^{(1)}$。

3. 第二级聚类

在距离矩阵 $D^{(1)}$的基础上，重复第一级聚类的做法，则得到第二级聚类的结果及第二级聚类后的距离矩阵 $D^{(2)}$。

4. 第 S 级聚类

在矩阵 $D^{(S-1)}$的基础上，仍重复第一级聚类的做法，可得第 S 级聚类结果及 $D^{(S)}$，直到全部样品(或变量)归并为一个大类为止。

在第 t 级聚类时，如果 $D^{(t-1)}$的非对角线元素中出现两个或两个以上最小值相等的元素，那么这些元素对应的类可同时聚为一类，也可以取其中的一类参加第 t 级聚类。

在聚合聚类中，若类 p 与 q 合并成新类 r，那么可由递推公式计算类 r 与某类 f 的距离，即

$$D_{rf}=\frac{1}{2}D_{fp}+\frac{1}{2}D_{fq}-\frac{1}{2}|D_{fp}-D_{fq}| \tag{4-14}$$

(二) 最长距离法

如果类 p 与 q 之间的距离定义为两类中相距最远的两个样品(或变量)点之间的距离，即

$$D_{pq}=\max_{\substack{X_i\in p\\X_j\in q}} d_{ij} \tag{4-15}$$

用这样定义的距离进行的聚合聚类叫做最长距离法。它的并类步骤与最短距离法的并类步骤完全相同，不同之处仅是：

若类 p 与 q 合并成新类 r 后，r 与某类 f 之间距离的递推公式为：

$$D_{rf}=\frac{1}{2}D_{fp}+\frac{1}{2}D_{fq}+\frac{1}{2}|D_{fp}-D_{fq}| \tag{4-16}$$

(三) 类平均法

类平均法定义两类之间的距离为样品(或变量)点之间距离的加权值，即

$$D_{pq}=\frac{1}{npnq}\sum_{\substack{X_i\in p\\x_j\in q}} d_{ij} \tag{4-17}$$

类 p 与 q 合并为类 r 后，计算它与某类 f 之间距离的递推公式为：

$$D_{rf}=\frac{np}{nr}D_{fp}+\frac{nq}{nr}D_{fq} \tag{4-18}$$

其中 nr、np、nq 分别为类 r、p、q 中的样品(或变量)数，且 $nr=np+nq$。

（四）重心法

设类 p、q 的重心分别是 $\overline{X}_p$ 和 $\overline{X}_q$，则类 p 与 q 之间的距离是

$$D_{pq}=d_{\bar{x}_p\bar{x}_q} \tag{4-19}$$

类 p 与 q 合并为类 r 后，它的重心为

$$\overline{X}_r=\frac{1}{nr}(np\overline{X}_p+nq\overline{X}_q)$$

其中 np、nq 分别为类 p、q 内的样品（或变量）数，并且 $nr=np+nq$。

设某类 f 的重心为 $\overline{X}_f$，那么计算类 r 与 f 之间距离的递推公式为：

$$D_{rf}^2=\frac{np}{nr}D_{fp}^2+\frac{nq}{nr}D_{fq}^2-\frac{npnq}{n_r^2}D_{pq}^2 \tag{4-20}$$

用重心距离进行的聚类叫重心法。

二、相关系数和相似系数聚合法

对于相关系数和相似系数是对应于距离类统计量的前三种方法。聚类亦有近邻联接法、远邻联接法和类平均法。

设类 p 与 q 合并为类 r，那么计算类 r 与类 f 之间相关（或相似）度量的递推公式为：

（一）近邻联接法

$$R_{rf}=\frac{1}{2}R_{fp}+\frac{1}{2}R_{fq}+\frac{1}{2}|R_{fp}-R_{fq}| \tag{4-21}$$

（二）远邻联接法

$$R_{rf}=\frac{1}{2}R_{fp}+\frac{1}{2}R_{fq}-\frac{1}{2}|R_{fp}-R_{fq}| \tag{4-22}$$

（三）类平均法

$$R_{rf}=\frac{np}{nr}R_{fp}+\frac{nq}{nr}R_{fq} \tag{4-23}$$

三、聚类结果选择

在本节中共介绍了七种聚合聚类方法。这些方法在计算步骤上是完全一样的，不同之处仅仅是类间聚类统计量的定义。但是，采用不同的聚类方法，分类的结果并不完全一致，如云南某地区超基性岩体岩样的聚类，如图 4-2 所示。究竟哪一种方法的分类结果好呢？目前尚无一个合适的衡量标准。在实际应用中，要结合其他地质理论及资料，进一步分析不同方法给出的分类结果，从中选择一种认为是合理的分类方案。

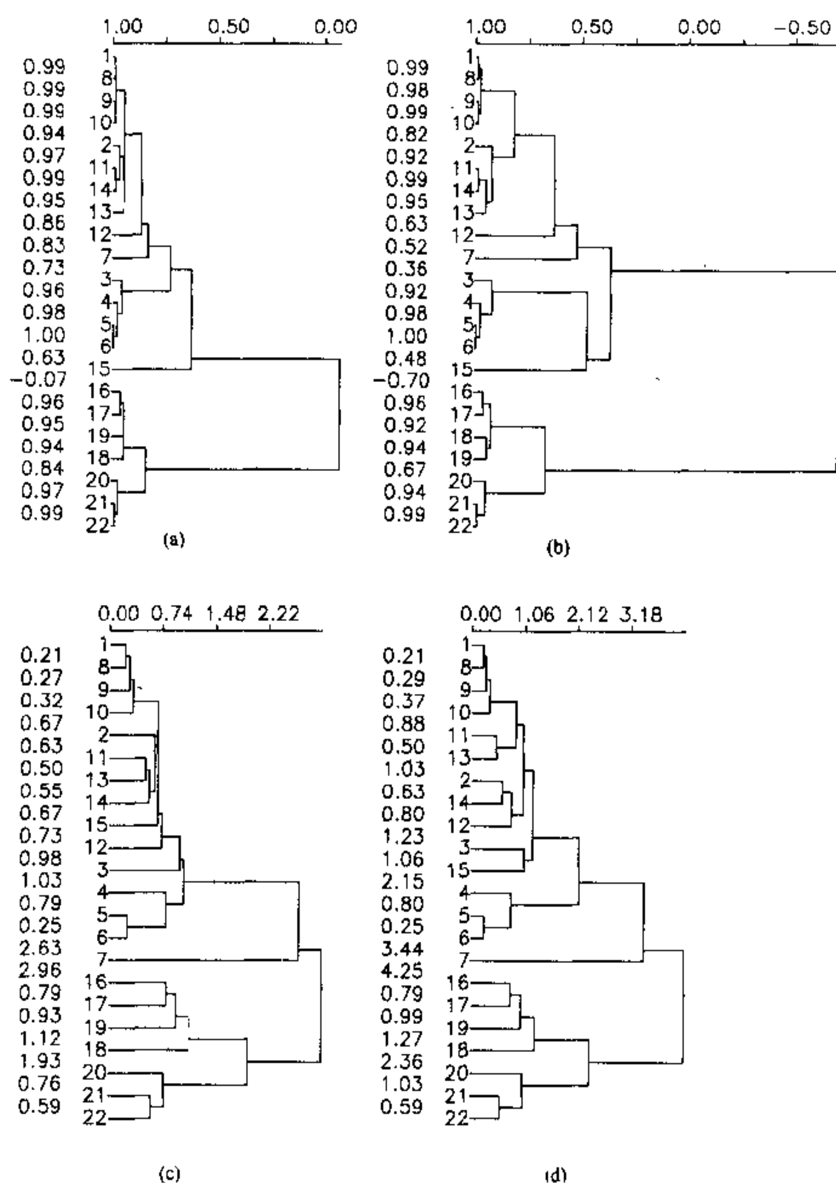


图 4-2 超基性岩体岩样聚类谱系图

(a) 相关系数近邻联接法 (b) 相关系数类平均法

(c) 欧氏距离最短距离法 (d) 欧氏距离类平均法

四、数据预处理

由于样品中各个变量的单位和数量级可能是不一样的，既便有些变量的度量相同，但各变量的绝对值大小也不相同，若直接用变量的原始观测值进行计算就会突出绝对值大的变量，而压低那些绝对值小的变量的作用，因此，在进行聚类分析时，应首先对原始数据进行预处理(见第二章)。

§3 有序样品聚类分析——最优分割法

一、最优分割

如果每个样品有 m 个变量，那么 n 个样品变量的观测值按样品相邻的顺序排列起来，得数据矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

其中元素 x_{ij}表示第 i 个样品第 j 个变量的观测值。

矩阵 X 既是 n 个有序样品。又是一个有序数列。这里所说的分割，是指在不打乱样品相邻关系的条件下，把 n 个有序样品进行分组(类、段)，也就是在不改变 X 中元素排列顺序的条件下，把有序数列进行分段。

把 n 个有序样品分为 k 组，可有 C_{n-1}^{k-1}种分法。其中，把 n 个有序样品分为 k 组后，使得各组内样品的差异最小，而各组之间样品的差异为最大的分法称为最优 k 分割法，相应的分割结果为最优 k 分割。

组内样品差异的大小，可用与组内样品所对应的有序数列段内数据的变差来表示，比如第 i 到第 j 个样品所对应的有序数列段

$$\begin{Bmatrix} x_{i1} & x_{i2} & \cdots & x_{im} \\ x_{i+11} & x_{i+12} & \cdots & x_{i+1m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jm} \end{Bmatrix}$$

的变差可表示为

$$d_{ij} = \sum_{\alpha=1}^{j}\sum_{\beta=1}^{m}(x_{\alpha\beta} - \bar{x}_{\beta}(i,j))^2 \qquad (4\text{-}24)$$

$$(i,j=1,2,\cdots,n)$$

其中

$$\bar{x}_{\beta}(i,j) = \frac{1}{j-i+1}\sum_{\alpha=i}^{j}x_{\alpha\beta} \qquad (4\text{-}25)$$

$$(\beta = 1,2,\cdots,m)$$

d_{ij}表示有序数列段内数据的变化情况，d_{ij}越小，有序数列段内数据的变化越小，即样品段内样品的性质越相近；反之，样品之间的差异就越明显。最优分割要求各段内部的差异达到最小，即各段变差的总和——段内离差平方和为最小。可以证明，任何一个有序数列的总离差平方和 S 由段内离差平方和 S_1 和段间离差平方和 S_2 两部分组成。证明如下：

如果把有序样品分为 k 组，每组内有 $n_l(l=1,2,\cdots,k)$个样品，相应的有序数列段为

$$\text{第一段：}\begin{Bmatrix} x_{11}^{(1)} & x_{11}^{(2)} & \cdots & x_{11}^{(m)} \\ x_{12}^{(1)} & x_{12}^{(2)} & \cdots & x_{12}^{(m)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{1n1}^{(1)} & x_{1n1}^{(2)} & \cdots & x_{1n1}^{(m)} \end{Bmatrix}$$

$$
\text{第二段：}\left\{\begin{matrix} x_{21}^{(1)} & x_{21}^{(2)} & \cdots & x_{21}^{(m)} \\ x_{22}^{(1)} & x_{22}^{(2)} & \cdots & x_{22}^{(m)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{2n_2}^{(1)} & x_{2n_2}^{(2)} & \cdots & x_{2n_2}^{(m)} \end{matrix}\right\}
$$

$$
\begin{matrix} \cdots\cdots & & \cdots\cdots \\ \cdots\cdots & & \cdots\cdots \\ \cdots\cdots & & \cdots\cdots \end{matrix}
$$

$$
\text{第}\,k\,\text{段：}\left\{\begin{matrix} x_{k1}^{(1)} & x_{k1}^{(2)} & \cdots & x_{k1}^{(m)} \\ x_{k2}^{(1)} & x_{k2}^{(2)} & \cdots & x_{k2}^{(m)} \\ \vdots & \vdots & \cdots & \vdots \\ x_{kn_k}^{(1)} & x_{kn_k}^{(2)} & \cdots & x_{kn_k}^{(m)} \end{matrix}\right\}
$$

那么，有序数列的总离差平方和

$$
\begin{aligned}
S &= \sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(x_{lj}^{(i)}-\overline{x}^{(i)})^2 = \sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}\left[(x_{lj}^{(i)}-x_l^{(i)})+(\overline{x}_l^{(i)}-\overline{x}^{(i)})\right]^2 \\
&= \sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(x_{lj}^{(i)}-\overline{x}_l^{(i)})^2+\sum_{l=1}^{k}\sum_{j=1}^{nl}\sum_{i=1}^{m}(\overline{x}_l^{(i)}-\overline{x}^{(i)})^2 \\
&\quad +2\sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(x_{lj}^{(i)}-\overline{x}_l^{(i)})(\overline{x}_l^{(i)}-\overline{x}^{(i)}) = S_1+S_2
\end{aligned}
$$

其中 $x_{lj}^{(i)}$ 是第 l 段内第 j 个样品第 i 个变量的观测值，并且：

$$
n=\sum_{l=1}^{k}n_l;\quad \overline{x}_l^{(i)}=\frac{1}{n_l}\sum_{j=1}^{n_l}x_{lj}^{(i)};\quad \overline{x}^{(i)}=\sum_{l=1}^{k}\sum_{j=1}^{n_l}x_{lj}^{(i)}\Big/\sum_{l=1}^{k}n_l=\sum_{l=1}^{k}n_lx_l^{(i)}/n;
$$

$$
2\sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(x_{lj}^{(i)}-\overline{x}_l^{(i)})(\overline{x}_l^{(i)}-\overline{x}^{(i)})
$$

$$
=2\sum_{l=1}^{k}\sum_{i=1}^{m}(\overline{x}_l^{(i)}-\overline{x}^{(i)})\sum_{j=1}^{n_l}(x_{lj}^{(i)}-\overline{x}_l^{(i)})=2\sum_{l=1}^{k}\sum_{i=1}^{m}(\overline{x}_l^{(i)}-\overline{x}^{(i)})(\overline{x}_l^{(i)}-\overline{x}_l^{(i)})n_l=0
$$

$$
S_1=\sum_{l=1}^{k}\sum_{j=1}^{n_l}\sum_{i=1}^{m}(x_{lj}^{(i)}-\overline{x}_l^{(i)})^2\ \text{为段内离差平方和；}
$$

$$
S_2=\sum_{l=1}^{k}n_l\sum_{i=1}^{m}(\overline{x}_l^{(i)}-\overline{x}^{(i)})^2\ \text{为段间离差平方和。}
$$

对于给定的有序数列，S 是个确定的值，因此，若使段内离差平方和 S_1 为最小，则段间离差平方和 S_2 必为最大。由此看来，使段内离差平方和为最小的分割法就是最优分割法。

二、段内变差的递推算法

前以叙及，段内离差平方和是段内变差的总和，这就意味着在最优分割的实现过程中，需要对段内变差进行大量的计算。由式(4-24)计算出的矩阵

$$
D=[d_{ij}]_{n\times n}
$$

称为段内变差矩阵，它是一个对称矩阵，且 $d_{ii}=0$。但是，当 n、m 较大时，其计算量还是相当大的，因此，采用一种改进的 D 矩阵计算方法。

由于

$$
d_{ij}=\sum_{\alpha=i}^{j}\sum_{\beta=1}^{m}(x_{\alpha\beta}-\overline{x}_{\beta}(i,j))^2
$$

$$= \sum_{\alpha=i}^{j} \sum_{\beta=1}^{m} \left(x_{\alpha\beta} - \frac{1}{j-i+1} \sum_{\alpha=i}^{j} x_{\alpha\beta} \right)^2$$

$$= \frac{1}{j-i+1} \sum_{\beta=1}^{m} \left[(j-i+1) \sum_{\alpha=i}^{j} x_{\alpha\beta}^2 - \left(\sum_{\alpha=i}^{j} x_{\alpha\beta} \right)^2 \right]$$

记

$$d'_{ij} = (j-i+1) \sum_{\alpha=i}^{j} x_{\alpha\beta}^2 - \left(\sum_{\alpha=i}^{j} x_{\alpha\beta} \right)^2$$

则有

$$\begin{aligned} d'_{ij} = & (j-i+1)(x_{i\beta}^2 + x_{i+1,\beta}^2 + \cdots + x_{j\beta}^2) - 2x_{i\beta}(x_{i+1,\beta} + x_{i+2,\beta} + \cdots + x_{j\beta}) \\ & - 2x_{i+1,\beta}(x_{i+2,\beta} + x_{i+3,\beta} + \cdots + x_{j\beta}) - \cdots - 2x_{j-2,\beta}(x_{j-1,\beta} + x_{j\beta}) \\ & - 2x_{j-1,\beta}x_{j\beta} \qquad (j > i) \end{aligned}$$

进而有

$$\begin{aligned} d'_{ij} = & (x_{i\beta} - x_{i+1,\beta})^2 + (x_{i\beta} - x_{i+2,\beta})^2 + \cdots + (x_{i\beta} - x_{j\beta})^2 \\ & + (x_{i+1,\beta} - x_{i+2,\beta})^2 + (x_{i+1,\beta} - x_{i+3,\beta})^2 + \cdots + (x_{i+1,\beta} - x_{j\beta})^2 \\ & + \cdots + (x_{j-2,\beta} - x_{j-1,\beta})^2 + (x_{j-2,\beta} - x_{j\beta})^2 + (x_{j-1,\beta} - x_{j\beta})^2 \\ = & \sum_{a=i}^{j-1} \sum_{b=a+1}^{j} (x_{a\beta} - x_{b\beta})^2 \qquad (j > i) \end{aligned}$$

因此

$$d_{ij} = \frac{1}{j-i+1} \sum_{a=i}^{j-1} \sum_{b=a+1}^{j} \sum_{\beta=1}^{m} (x_{a\beta} - x_{b\beta})^2 \quad (j > i)$$

令

$$d''_{ij} = \sum_{a=i}^{j-1} \sum_{b=a+1}^{j} \sum_{\beta=1}^{m} (x_{a\beta} - x_{b\beta})^2 \quad (j > i) \tag{4-26}$$

并称其为未加权的段内变差。再令

$$y_{ij} = \sum_{\beta=1}^{m} (x_{i\beta} - x_{j\beta})^2 \qquad (i = 1,2,\cdots,n; \quad j = i,i+1,\cdots,n)$$

则 y_{ij}组成一个上三角矩阵，即

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ & y_{22} & \cdots & \vdots \\ & & \ddots & \vdots \\ & & & y_{nn} \end{bmatrix}$$

同时，检查由式(4-26)得来的 $d''_{ij}(j>i)$加上 d''_{ii}(等于 0)组成的上三角矩阵为

$$D'' = \begin{bmatrix} d''_{11} & d''_{12} & \cdots & d''_{1n} \\ & d''_{22} & \cdots & \vdots \\ & & \ddots & \vdots \\ & & & d''_{nn} \end{bmatrix}$$

由此看出，d''_{ij}实质上等于 Y 矩阵中以 y_{ii}、y_{jj}、y_{ij}为顶点的三角形中全部元素之和。在实际计算中，若对 D''矩阵按列从左到右，每列中从下向上计算，并令

$$r = \sum_{k=i+1}^{j} y_{kj}$$

则有

$$d''_{ij}=d''_{i-1,j-1}+y_{ij}+r$$

最终得

$$\begin{cases} d_{ij}=\dfrac{1}{j-i+1}d''_{ij} \\ d''_{ij}=d''_{i-1,j-1}+y_{ij}+r \end{cases} \qquad (i,j=1,2,\cdots,n) \tag{4-27}$$

式(4-27)是计算段内变差的递推算法。当 n、m 较大时，计算效果显著。

三、如何实现最优分割

为便于书写，把 n 个有序样品改记为如下形式：

$$X=[X_1X_2\cdots X_n]'$$

其中

$$X_i=(x_{i1}\ x_{i2}\cdots\ x_{im}) \qquad (i=1,2,\cdots,n)$$

现分别讨论对 X 实现最优二分割、最优三分割，…最优 k 分割的方法。

1. 最优二分割

记二分割的段内离差平方和为 $S_n(2;j)$，其中 n 是参加分割的样品数；2 表示把 n 个有序样品分为二段；j 表示以第 $j(1\leqslant j\leqslant n-1)$ 个样品后为分割点。

由段内变差矩阵 D，可求得 n 个样品分为二段的段内离差平方和，即

$$S_n(2;j)=d_{1j}+d_{j+1,n} \qquad (j=1,2,\cdots,n-1) \tag{4-28}$$

然后从中找出一个最小值，即

$$S_n(2;\alpha_1(n))=\min_{1\leqslant j\leqslant n-1}S_n(2;j)$$

这个最小值就是 n 个样品最优二分割的段内离差平方和，它所对应的分点 $\alpha_1(n)$ 为 n 个样品最最优二分割的分割点。分割结果为

$$[(X_1X_2\cdots X_{\alpha_1(n)})(X_{\alpha_1(n)+1}\cdots X_n)]'$$

2. 最优三分割

若记三分割的段内离差平方和为 $S_n(3;\alpha_1(j),j)$，根据段内离差平方和的算法，那么有

$$S_n(3;\alpha_1(j),j)=d_{1\alpha_1(j)}+d_{\alpha_1(j)+1,j}+d_{j+1,n}=S_j(2;\alpha_1(j))+d_{j+1,n}$$

$$(j=2,3,\cdots,n-1)$$

其中 $\alpha_1(j)$、j 为三分割的两个分割点，且 $1\leqslant\alpha_1(j)\leqslant j-1$。

如果 $S_n(3;\alpha_1(j),j)$ 为最优三分割，那么 $S_j(2;\alpha_1(j))$ 必为最优二分割。否则，必存在另外一个最优二分割 $S_j(2;\alpha'_1(j))$，使得

$$S_n(3;\alpha_1(j),j)>S_n(3;\alpha'_1(j),j)$$

这与 $S_n(3;\alpha_1(j),j)$ 是最优三分割相矛盾，因此，在对 n 个样品进行最优三分割时，必须先求出前 $j(j=2,3,\cdots,n-1)$ 个样品的最优二分割，得到分割点 $\alpha_1(j)$，由

$$[(X_1X_2\cdots X_{\alpha_1(j)})(X_{\alpha_1(j)-1}\cdots X_j)]'$$

与 $(X_{j+1},\cdots X_n)'$ 构成一个三分割，然后找出一个合适的 j 使得

$$S_n(3;\alpha_1(j),j)=S_j(2;\alpha_1(j))+d_{j+1,n}$$

尽可能的小。若

$$S_n(3;\alpha_1(j),\alpha_2(j))=\min_{1\leqslant j\leqslant n-1}S_n(3;\alpha_1(j),j)$$

那么 $\alpha_1(j)$、$\alpha_2(j)$ 为 n 个样品的最优三分割点，$\alpha_1(j)$ 为前 j 个样品最优二分割点。最优三分割结果为

$$[(X_1X_2\cdots X_{\alpha_1(j)})(X_{\alpha_1(j)+1}\cdots X_{\alpha_2(j)})(X_{\alpha_2(j)+1}\cdots X_n)]'$$

综上所述，得出三分割中各种分割相应的段内离差平方和为

$$S_p(3;\alpha_1(j),j)=S_j(2;\alpha_1(j))+d_{j+1,p} \tag{4-29}$$

$$(p=3,4,\cdots,n;\quad j=2,3,\cdots,p-1)$$

从中找出一个最小值，即

$$S_p(3;\alpha_1(p),\alpha_2(p))=\min_{2\leqslant j\leqslant p-1}S_p(3;\alpha_1(j),j)$$

从而得到前 p 个样品的最优三分割

$$[(X_1X_2\cdots X_{\alpha_1(p)})(X_{\alpha_1(p)+1}\cdots X_{\alpha_2(p)})(X_{\alpha_2(p)+1}\cdots X_p)]'$$

3. 最优 k 分割

在已作出最优 $k-1$ 分割后，分别计算

$$\begin{aligned}&S_p(k;\alpha_1(j),\alpha_2(j),\cdots,\alpha_{k-2}(j),j)\\&=S_j(k-1;\alpha_1(j),\alpha_2(j),\cdots,\alpha_{k-2}(j))+d_{j+1,p}\end{aligned} \tag{4-30}$$

$$(p=k,k+1,\cdots,n;\quad j=k-1,k,\cdots,p-1)$$

并从中找出一个最小值，即

$$S_p(k;\alpha_1(p),\alpha_2(p),\cdots,\alpha_{k-2}(p),\alpha_{k-1}(p))=\min_{k-1\leqslant j\leqslant p-1}S_p(k;\alpha_1(j),\alpha_2(j),\cdots\alpha_{k-2}(j),j)$$

其中 $\alpha_1(p),\alpha_2(p),\cdots,\alpha_{k-2}(p)$ 为前 p 个样品的最优 $k-1$ 分割的分割点，并与 $\alpha_{k-1}(p)$ 一起构成前 p 个样品的最优 k 分割点。

n 个样品最优 k 分割的段内离差平方和为

$S_n(k;\alpha_1(n),\alpha_2(n),\cdots,\alpha_{k-1}(n))$，分割点为 $\alpha_1(n),\alpha_2(n),\cdots,\alpha_{k-1}(n)$。

4. 分割段数的确定

对于分割段数 k 一直可以做到预先给定的整数 k 为止，或者预先给定一个小正数 δ，使得段内离差平方和

$$S_n(k;\alpha_1(n),\alpha_2(n),\cdots,\alpha_{k-1}(n))<\delta$$

后就不再继续分割了，这样得出的 k 就是最后分割的段数。由图 4-3 可见，段内离差平方和 S_n 将随着分段数 k 的增加而单调递减，并趋于平缓，因此，可以选择开始趋于平缓时的分割数为最优分割数。

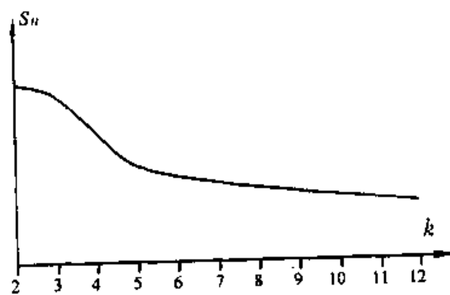


图 4-3 S_n 与 k 变化关系示意图

四、最优分割的具体计算步骤

1. 数据正规化

为了消除样品中不同变量观测值数量级的差别对分割结果的影响，要对原始数据进行处理。

设原始数据为 $X=[x_{ij}]_{n\times m}$，把它的元素 x_{ij} 按

$$z_{ij}=\frac{x_{ij}-\min\limits_{1\leqslant i\leqslant n}\{x_{ij}\}}{\max\limits_{1\leqslant i\leqslant n}\{x_{ij}\}-\min\limits_{1\leqslant i\leqslant n}\{x_{ij}\}}\qquad(i=1,2,\cdots,n;j=1,2,\cdots,m)$$

进行变换，得新的数据 $Z=[z_{ij}]_{n\times m}$。

2. 计算段内变差矩阵

按下式

$$d_{ij}=\sum_{\alpha=1}^{j}\sum_{\beta=1}^{m}[z_{\alpha\beta}-\bar{z}_{\beta}(i,j)]^2$$

$$(i,j=1,2,\cdots,n)$$

计算段内变差矩阵 $D=[d_{ij}]_{n\times n}$。式中

$$\bar{z}_{\beta}(i,j)=\frac{1}{j-i+1}\sum_{\alpha=i}^{j}z_{\alpha\beta}\quad(\beta=1,2\cdots,m)$$

3. 最优二分割

由段内变差矩阵 D，按式(4-28)计算二分割中各种分割相应的段内离差平方和，从中找出一个最小值，即

$$S_p(2;\alpha_1(p))=\min_{k-1\leqslant j\leqslant p-1}S_p(2;j)$$

从而确定前 p 个样品的最优二分割。

4. 最优三分割

按式(4-29)及 D 矩阵计算三分割中不同分割相应的段内离差平方和，从中找出一个最小值，即

$$S_p(3;\alpha_1(p)\alpha_2(p)=\min_{2\leqslant j\leqslant p-1}S_p(3;\alpha_1(j),j)$$

确定三分割的分割点，得到前 p 个样品的最优三分割。

5. 最优 k 分割

由式(4-30)及 D 矩阵计算 k 分割中各种分割相应的段内离差平方和，从中找出一个最小值，即

$$S_p(k;\alpha_1(p),\alpha_2(p),\cdots,\alpha_{k-2}(p),\alpha_{k-1}(p))=\min_{k-1\leqslant j\leqslant p-1}S_p(k;\alpha_1(j),\alpha_2(j),\cdots,\alpha_{k-2}(j),j)$$

确定 k 分割的分割点，从而得到前 p 个样品的最优 k 分割。

6. 最优分割示例

为了进一步理解最优分割的计算步骤及段内离差平方和的计算方法，下面给出一个最优四分割的示意性例子。

(1) 原始数据

表 4-1 是某井段自然伽玛、自然电位、电阻率测井的六组实测值及正规化数据。

表 4-1　某井段测井数据

测值深度/m	自然伽玛		自然电位		电阻率	
	实测值	正规化值	实测值	正规化值	实测值	正规化值
1885	52.99	0.0000	2.76	0.7108	94.85	1.0000
1886	53.36	0.0058	2.83	0.7952	77.22	0.8078
1887	73.36	0.3170	2.39	0.2651	60.61	0.6267
1888	78.58	0.3982	2.17	0.0000	35.77	0.3559
1889	77.24	0.3774	3.00	1.0000	11.55	0.0918
1890	117.25	1.0000	2.22	0.0602	3.13	0.0000

(2) 段内变差矩阵

根据式(4-24)，由表 4-1 正规化数据计算段内变差矩阵的下三角矩阵

$$D=\begin{bmatrix} 0.0000 \\ 0.0220 & 0.0000 \\ 0.2978 & 0.2053 & 0.0000 \\ 0.7782 & 0.5172 & 0.0751 & 0.0000 \\ 1.3513 & 1.0361 & 0.6834 & 0.5351 & 0.0000 \\ 2.3390 & 1.7996 & 1.1808 & 0.9473 & 0.6396 & 0.0000 \end{bmatrix}$$

(3) 最优二分割

对 $p=2,3,\cdots,6$ 计算 $S_p(2;j)=d_{1j}+d_{j+1,p}$ $(j=1,2,\cdots,p-1)$

当 $p=2$ 时：

$S_2(2;1)=d_{11}+d_{22}=0$

$S_2(2;\alpha_1(2))=\min\limits_{1\leqslant j\leqslant 1} S_2(2;j)=S_2(2;1)=0$， $\alpha_1(2)=1$

当 $p=3$ 时：

$S_3(2;1)=d_{11}+d_{23}=0.2053$ $S_3(2;2)=d_{12}+d_{33}=0.0220$

$S_3(2;\alpha_1(3))=\min\limits_{1\leqslant j\leqslant 2} S_3(2;j)=S_3(2;2)=0.0220$， $\alpha_1(3)=2$

当 $p=4$ 时：

$S_4(2;1)=d_{11}+d_{24}=0.5172$

$S_4(2;2)=d_{12}+d_{34}=0.0971$

$S_4(2;3)=d_{13}+d_{44}=0.2978$

$S_4(2;\alpha_1(4))=\min\limits_{1\leqslant j\leqslant 3} S_4(2;j)=S_4(2;2)=0.0971$， $\alpha_1(4)=2$

当 $p=5$ 时：

$S_5(2;1)=d_{11}+d_{25}=1.0361$

$S_5(2;2)=d_{12}+d_{35}=0.7054$

$S_5(2;3)=d_{13}+d_{45}=0.8329$，

$S_5(2;4)=d_{14}+d_{55}=0.7782$

$S_5(2;\alpha_1(5))=\min\limits_{1\leqslant j\leqslant 4} S_5(2;j)=S_5(2;2)=0.7054$， $\alpha_1(5)=2$

当 $p=6$ 时：

$S_6(2;1)=d_{11}+d_{26}=1.7996$

$S_6(2;2)=d_{12}+d_{36}=1.2028$

$S_6(2;3)=d_{13}+d_{46}=1.2451$

$S_6(2;4)=d_{14}+d_{56}=1.4178$

$S_6(2;5)=d_{15}+d_{66}=1.3513$

$S_6(2;\alpha_1(6))=\min\limits_{1\leqslant j\leqslant 5} S_6(2;j)=S_6(2;2)=1.2028$， $\alpha_1(6)=2$

由以上结果可知，六个样品最优二分割的分割点为 2，分割结果为

$$[(X_1X_2)(X_3X_4X_5X_6)]'。$$

(4) 最优三分割

对 $p=3,4,5,6$ 计算 $S_p(3;\alpha_1(j),j)=S_j(2;\alpha_1(j))+d_{j+1,p}$ $(j=2,3,\cdots,p-1)$

当 $p=3$ 时：

$S_3(3;\alpha_1(2),2)=S_2(2;\alpha_1(2))+d_{33}=0$

$S_3(3;\alpha_1(3),\alpha_2(3))=\min_{2\leqslant j\leqslant 2}S_3(3;\alpha_1(j),j)=S_3(3;\alpha_1(2),2)=0$

$\alpha_1(3)=\alpha_1(2)=1,\alpha_2(3)=2$

当 $p=4$ 时：

$S_4(3;\alpha_1(2),2)=S_2(2;\alpha_1(2))+d_{34}=0.0751$

$S_4(3;\alpha_1(3),3)=S_3(2;\alpha_1(3))+d_{44}=0.0220$

$S_4(3;\alpha_1(4),\alpha_2(4))=\min_{2\leqslant j\leqslant 3}S_4(3;\alpha_1(j),j)=S_4(3;\alpha_1(3),3)=0.0220$

$\alpha_1(4)=\alpha_1(3)=2,\quad \alpha_2(4)=3$

当 $p=5$ 时：

$S_5(3;\alpha_1(2),2)=S_2(2;\alpha_1(2))+d_{35}=0.6834$

$S_5(3;\alpha_1(3),3)=S_3(2;\alpha_1(3))+d_{45}=0.5571$

$S_5(3;\alpha_1(4),4)=S_4(2;\alpha_1(4))+d_{55}=0.0971$

$S_5(3;\alpha_1(5),\alpha_2(5))=\min_{2\leqslant j\leqslant 4}S_5(3;\alpha_1(j),j)=S_5(3;\alpha_1(4),4)=0.0971,$

$\alpha_1(5)=\alpha_1(4)=2,\quad \alpha_2(5)=4$

当 $p=6$ 时：

$S_6(3;\alpha_1(2),2)=S_2(2;\alpha_1(2))+d_{36}=1.1808$

$S_6(3;\alpha_1(3),3)=S_3(2;\alpha_1(3))+d_{46}=0.9693$

$S_6(3;\alpha_1(4),4)=S_4(2;\alpha_1(4))+d_{36}=0.7367$

$S_6(3;\alpha_1(5),5)=S_5(2;\alpha_1(5))+d_{66}=0.7054$

$S_6(3;\alpha_1(6),\alpha_2(6))=\min_{2\leqslant j\leqslant 5}S_6(3;\alpha_1(j),j)=S_6(3;\alpha_1(5),5)=0.7054$

$\alpha_1(6)=\alpha_1(5)=2,\quad \alpha_2(6)=5$

六个样品的最优三分割点为 2、5，分割结果为$[(X_1X_2)(X_3X_4X_5)(X_6)]'$。

(5) 最优四分割

对 $p=4,5,6$ 计算 $S_p(4;\alpha_1(j),\alpha_2(j),j)=S_j(3;\alpha_1(j),\alpha_2(j))+d_{j+1,p}$

$(j=3,4,5,\cdots,p-1)$

当 $p=4$ 时：

$S_4(4;\alpha_1(3),\alpha_2(3),3)=S_3(3;\alpha_1(3),\alpha_2(3))+d_{44}=0.0000$

$S_4(4;\alpha_1(4),\alpha_2(4),\alpha_3(4))=\min_{3\leqslant j\leqslant 3}S_4(4;\alpha_1(j),\alpha_2(j),j)=S_4(4;\alpha_1(3),\alpha_2(3),3)=0$

$\alpha_1(4)=\alpha_1(3)=1,\alpha_2(4)=\alpha_2(3)=2,\alpha_3(4)=3$

当 $p=5$ 时：

$S_5(4;\alpha_1(3),\alpha_2(3),3)=S_3(3;\alpha_1(3),\alpha_2(3))+d_{45}=0.5351$

$S_5(4;\alpha_1(4),\alpha_2(4),4)=S_4(3;\alpha_1(4),\alpha_2(4))+d_{55}=0.0220$

$S_5(4;\alpha_1(5),\alpha_2(5),\alpha_3(5))=\min_{3\leqslant j\leqslant 4}S_5(4;\alpha_1(j),\alpha_2(j),j)$

$=S_5(4;\alpha_1(4),\alpha_2(4),4)=0.0220$

$\alpha_1(5)=\alpha_1(4)=2,\alpha_2(5)=\alpha_2(4)=3,\alpha_3(5)=4$

当 $p=6$ 时：

$S_6(4;\alpha_1(3),\alpha_2(3),3)=S_3(3;\alpha_1(3),\alpha_2(3))+d_{46}=0.9473$

$S_6(4;\alpha_1(4),\alpha_2(4),4)=S_4(3;\alpha_1(4),\alpha_2(4))+d_{56}=0.6396$

$S_6(4;\alpha_1(5),\alpha_2(5),5)=S_5(3;\alpha_1(5),\alpha_2(5))+d_{66}=0.0971$

$$S_6(4;\alpha_1(6),\alpha_2(6),\alpha_3(6))=\min_{3\leqslant j\leqslant 5} S_6(4;\alpha_1(j),\alpha_2(j),j)$$
$$=S_6(4;\alpha_1(5),\alpha_2(5),5)=0.0971$$

$\alpha_1(6)=\alpha_1(5)=2,\alpha_2(6)=\alpha_2(5)=4,\alpha_3(6)=5$

六个样品最优四分割的分割点为2、4、5，分割结果为$[(X_1X_2)(X_3X_4)(X_5)(X_6)]'$。

§4 聚类分析FORTRAN源程序

一、聚合法聚类分析程序

程序用于对具有m个变量的n个样品，选择一种或几种聚类统计量，用聚合法进行Q型或R型聚类分析，并绘制聚类谱系图。现将程序中的主要参数、符号及程序使用方法说明如下：

（一）主要参数及符号

1. 参数

n——整型变量，样品数；

m——整型变量，每个样品的变量数；

x0、y0——绘图坐标原点；

dh——轴字符大小控制变量；

ii——是否输出相似性度量的整型变量。

2. 符号

xm——原始数据数组名；

x——预处理后的数据数组名；

rq——存放相似性度量的数组名；

ns——数据预处理类型选择变量；

nd——相似性度量选择变量；

nc——系统聚类法选择变量。

3. 子程序

jhjl——聚合聚类子程序；

norm——数据极差标准化子程序；

stand——数据标准差标准化子程序；

cor——计算相关系数子程序；

cosin——计算相似系数子程序；

eds——计算欧氏距离系数子程序；

obdos——计算斜交距离系数子程序；

cp1——挑选最大的相似性度量（相关系数、相似系数）的子程序；

cp2——挑选最小的相似性度量（距离系数）子程序；

cd1——逐步计算、修改相关（或相似）系数子程序；

cd2——逐步计算、修改距离系数子程序；

flt——计算聚类谱系图上各线段起点和终点坐标子程序；

plot——绘制聚类谱系图子程序。

（二）程序使用说明

1. 数据文件

数据文件由 $n\times m$ 个样品观测值构成，其格式为：

$$
\begin{array}{cccc}
x_{11}, & x_{12}, & \cdots, & x_{1m} \\
x_{21}, & x_{22}, & \cdots, & x_{2m} \\
\vdots & & & \vdots \\
x_{n1}, & x_{n2}, & \cdots, & x_{nm}
\end{array}
$$

2. 操作说明

在 DOS 操作系统下，键入聚类分析目标程序名 jlfx 具体操作步骤如下：

(1) 输入数据文件名(Input your data file name)；

(2) 输入样品数 n、变量数 m、程控制变量 nm(input n,m,nm)；

(3) 选择数据预处理类型(To select a data pre—processing mode)。ns=0，数据不进行预处理；ns=1，表示要对原始数据作标准差标准化；ns=2，对原始数据作极差标准化预处理。

(4) 决定聚类类型(To select a cluster mode)选择 Q 或 r 表示 Q 型或 R 型聚类分析。

(5) 选择相似性度量(To select a simil metric)。输入 nd(nd=1,2,3,4，分别表示相关系数、相似系数、欧氏距离和斜交距离)值。

(6) 选择系统聚类法(To select a hierachical clustering method)。输入 nc(nc=1,2,3，分别为近邻联接法、远邻联接法和类平均法)的值。

第一种聚类方法结束后，屏幕显示：

Another method (y/n)?

回答 y，表示换一种聚类方法，程序返回到 nc 选择处；键入 n，同样出现 nc 选择，nc=1,2,3,4 分别为最短距离法、最长距离法、类平均法和重心法。

(7) 最后一次聚类结束，将先后显示：

Another simitarity metric (y/n)?

Another cluster mode (y/n)?

Another data pre—processing mode (y/n)?

键入 y，将分别返回到选择相似性度量、聚类类型和数据预处理选择处；若键入 n，程序运行结束。

3. 主要输出结果

输出选定相似性度量和聚合法聚类分析的聚类谱系图。

(三) 源程序

1. 聚类分析流程

聚类分析流程如图 4-4 所示。

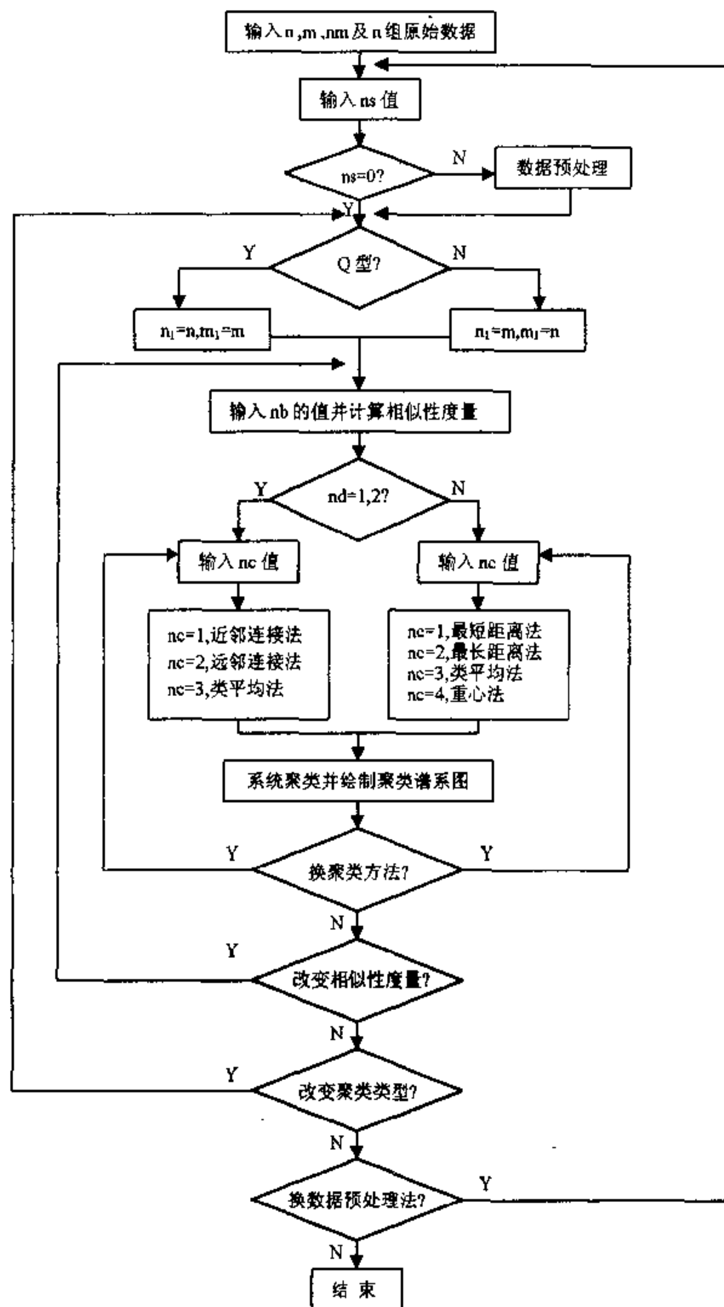


图 4-4　聚类分析流程图

2. 聚类分析 FORTRAN 源程序

```
c          program—jlfx. for
$large
$debug    program jlfx
          dimension xm(150,150)
```

```
      common /og/x(150,150),rq(150,150),nn(150)
      common /ofl/r(150),ib2(150)
      character fname * 10,type
      external cp1,cp2,cd1,cd2,cd3
      write( * . * ) 'Please input your data file name '
      read( * ,'(a)') fname
      write( * , * ) 'Please input n,m,nm '
      read( * , * ) n,m,nm
      open(3,file=fname)
      read(3, * ) ((xm(i,j),j=1,m),i=1,n)
10    write( * , * ) 'To select a data pre-processing mode'
      do 20 i=1,n
      do 20 j=1,m
20    x(i,j)=xm(i,j)
      if(nm. eq. 0) write( * ,30) ((x(i,j),j=1,m),i=1,n)
30    format(1x,6f10. 4)
      write( * ,'(a $ )') ' ns:0,1,2,ns='
      read( * , * ) ns
      if(ns. eq. 1) call stand(n,m)
      if(ns. eq. 2) call norm(n,m)
40    write( * , * ) ' To select acluster mode:q or r '
      read( * ,'(a)') type
      if(type. eq. 'R'. or. type. eq. 'r') then
      do 50 i=1,n
      do 50 j=1,m
      rq(j,i)=x(i,j)
50    continue
      do 60 i=1,m
      do 60 j=1,n
      x(i,j)=rq(i,j)
60    continue
      n1=m
      m1=n
      else
      n1=n
      m1=m
      end if
70    write( * . * )' To select a similarity metric'
      write( * ,'(a $ )') ' nd:1,2,3,4,nd= '
      read( * , * ) nd
```

```
      if(nd.eq.1) call cor(n1,m1)
      if(nd.eq.2) call cosin(n1,m1)
      if(nd.eq.3) call eds(n1,m1)
      if(nd.eq.4) call obds(n1,m1)
      if(nd.eq.1.or.nd.eq.2) then
80    write(*,'(2a/a$)') ' To select a hierachical',
     # ' lustering method ',' nc:1,2,3,nc= '
      read(*,*) nc
      call jhjlf(n1,nc,1,cp1,cd1)
      write(*,'(//a)') ' Another method (y/n) '
      read(*,'(a)')type
      if(type.eq.'Y'.or.type.eq.'y') then
      do 90 i=1,n1-1
      do 90 j=i+1,n1
      rq(i,j)=rq(j,i)
90    continue
      go to 80
      end if
      else
95    write(*,'(2a/a$)') ' To select a bierachical clustering',
     # ' method',' nc:1,2,3,4, nc='
      read(*,*) nc
      call jhjlf(n1,nc,2,cp2,cd2)
      write(*,'(//a)') ' Another method (y/n) ? '
      read(*,'(a)') type
      if(type.eq.'Y'.or.type.eq.'y') then
      do 99 i=1,n1-1
      do 99 j=i+1,n1
99    rq(i,j)=rq(j,i)
      go to 95
      end if
      end if
      write(*,'(///a)') ' Another similarity metric (y/n) '
      read(*,'(a)') type
      if(type.eq.'Y'.or.type.eq.'y') go to 70
      write(*,*) 'Another cluster mode (y/n) '
      read(*,'(a)') type
      if(type.eq.'Y'.or.type.eq.'y') go to 40
      write(*,*) ' Another data pre-processing mode (y/n) '
      read(*,'(a)') type
```

```
      if(type.eq.'Y'.or.type.eq.'y') go to 10
      stop
      end

      subroutine jhjlf(n,nc,nt,cp,cd)
      dimension a(150),mb(150,2),ns(150),l(150)
      common /og/x(150,150),rd(150,150),nn(150)
      common /ofl/r(150),ib2(150)
      external cp,cd
      do 20 i=1,n
      l(i)=i
20    nn(i)=1
      do 60 k=1,n-1
      call cp(n,k,ii,jj,rdm)
      a(k)=rdm
      if(l(ii).gt.l(jj)) then
      lj=l(ii)
      l(ii)=l(jj)
      l(jj)=lj
      do 25 i=1,ii-1
      c1=rd(i,ii)
      rd(i,ii)=rd(i,jj)
25    rd(i,jj)=c1
      if(ii.eq.jj-1) go to 35
      do 30 i=ii+1,jj-1
      c1=rd(ii,i)
      rd(ii,i)=rd(i,jj)
30    rd(i,jj)=c1
      end if
35    mb(k,1)=l(ii)
      mb(k,2)=l(jj)
      nn(ii)=nn(ii)+nn(jj)
      if(jj.ne.n+1-k) then
      nn1=nn(jj)
      nn(jj)=nn(n+1-k)
      nn(n+1-k)=nn1
      lj=l(jj)
      l(jj)=l(n+1-k)
      l(n+1-k)=lj
      do 40 i=1,jj-1
```

```
      c1=rd(i,jj)
      rd(i,jj)=rd(i,n+1-k)
40    rd(i,n+1-k)=c1
      if(jj.eq.n-k) go to 50
      do 45 i=jj+1,n-k
      c1=rd(jj,i)
      rd(jj,i)=rd(i,n+1-k)
45    rd(i,n+1-k)=c1
      end if
50    do 55 i=1,n-k
      if(i.ne.ii) call cd(n,k,ii,i,nc)
55    continue
60    continue
      do 65 i=1,n-1
65    ns(i)=0
      do 90 i=2,n-1
      do 70 j=1,i-1
      if(mb(i,2).eq.mb(j,1)) then
      ns(i)=j
      go to 75
      end if
70    continue
75    do 85 j=i-1,1,-1
      if(mb(i,1).eq.mb(j,1)) then
      l0=j
      do 80 k=1,n
      if(ns(l0).eq.0) then
      ns(l0)=i
      go to 90
      else
      l0=ns(l0)
      end if
80    continue
      end if
85    continue
90    continue
      do 92 i=1,n
92    l(i)=0
      do 94 i=1,n-1
      if(ns(i).ne.0) l(ns(i))=1
```

```
94      continue
        do 96 i=1,n
        if(l(i).eq.0) then
        ii=i
        go to 98
        end if
96      continue
98      r(1)=0
        r(2)=a(ii)
        ib2(1)=mb(ii,1)
        ib2(2)=mb(ii,2)
        do 99 k=1,n-2
        ib2(k+2)=mb(ns(ii),2)
        r(k+2)=a(ns(ii))
        ii=ns(ii)
99      continue
        call flt(n,nt)
        call plot(n,nt)
        return
        end

        subroutine stand(n,m)
        common /og/x(150,150),rq(150,150),nn(150)
        do 30 j=1,m
        d=0.
        do 10 i=1,n
10      d=d+x(i,j)
        rq(j,1)=d/n
        d=0.
        do 20 i=1,n
20      d=d+(x(i,j)-rq(j,1))**2
        rq(j,2)=sqrt(d/(n-1))
        do 30 i=1,n
        x(i,j)=(x(i,j)-rq(j,1))/rq(j,2)
30      continue
        return
        end

        subroutine norm(n,m)
        common /og/x(150,150),rq(150,150),nn(150)
```

```
          do 30 j=1,m
          xmax=x(1,j)
          xmin=x(1,j)
          do 10 i=2,n
          if(x(i,j).gt.xmax) xmax=x(i,j)
          if(x(i,j).lt.xmin) xmin=x(i,j)
10        continue
          do 20 i=1,n
          x(i,j)=(x(i,j)-xmin)/(xmax-xmin)
20        continue
30        continue
          return
          end

          subroutine cor(n,m)
          common /og/x(150,150),r(150,150),nn(150)
          do 20 j=1,n
          d=0.
          do 10 =1,m
10        d=d+x(j,i)
20        r(j,1)=d/m
          do 40 i=1,n-1
          do 40 j=i+1,n
          di=0.
          dj=0.
          dij=0.
          do 30 k=1,m
          d1=x(i,k)-r(i,1)
          d2=x(j,k)-r(j,1)
          dij=dij+d1 * d2
          di=di+d1 * d1
30        dj=dj+d2 * d2
40        r(i,j)=dij/sqrt(di * dj)
          do 50 i=1,n
          r(i,i)=1
          if(i.eq.n) go to 60
          do 50 j=i+1,n
          r(j,i)=r(i,j)
50        continue
60        return
```

```
      end

      subroutine cosin(n,m)
      common /og/x(150,150),c(150,150),nn(150)
      do 20 i=1,n
      c(i,i)=1.
      if(i.eq.n) go to 30
      do 20 j=i+1,n
      di=0.
      dj=0.
      dij=0.
      do 10 k=1,m
      di=di+x(i,k)*x(i,k)
      dj=dj+x(j,k)*x(j,k)
10    dij=dij+x(i,k)*x(j,k)
      c(i,j)=dij/sqrt(di*dj)
20    c(j,i)=c(i,j)
30    continue
      return
      end

      subroutine eds(n,m)
      common /og/x(150,150),d(150,150),nn(150)
      do 20 i=1,n
      d(i,i)=0.
      if(i.eq.n) go to 30
      do 20 j=i+1,n
      dd=0.
      do 10 k=1,m
10    dd=dd+(x(i,k)-x(j,k))*(x(i,k)-x(j,k))
      d(i,j)=sqrt(dd)
20    d(j,i)=d(i,j)
30    return
      end

      subroutine obds(n,m)
      common /og/x(150,150),d(150,150),nn(150)
      dimension cx(150)
      do 10 i=1,n
      do 10 j=1,m
```

```
10          d(i,j)=x(i,j)
            do 20 i=1,m
            do 20 j=1,n
20          x(i,j)=d(j,i)
            call cor(m,n)
            n1=amax0(n,m)
            m1=amin0(n,m)
            do 50 i=1,m1
            do 30 j=i+1,n1
30          cx(j)=x(j,i)
            do 40 j=i+1,n1
40          x(j,i)=x(i,j)
            do 50 j=i+1,n1
50          x(i,j)=cx(j)
            do 70 i=1,n-1
            do 70 j=i+1,n
            d(i,j)=0
            do 60 l=1,m
            do 60 k=1,m
            if(k. ge. l) then
            r1=d(k,l)
            else
            r1=d(l,k)
            end if
            d1=(x(i,l)-x(j,l)) * (x(i,k)-x(j,k))
60          d(i,j)=d(i,j)+d1 * r1
            if(d(i,j). lt. 0. ) d(i,j)=-d(i,j)
70          d(i,j)=sqrt(d(i,j))
            do 80 i=1,n-1
            d(i,i)=0.
            do 80 j=i+1,n
            d(j,i)=d(i,j)
80          continue
            d(n,n)=0.
            return
            end

            subroutine cp1(n,k,ii,jj,rmax)
            common /og/x(150,150),r(150,150),nn(150)
            rmax=-1000.
```

```
      do 20 i=1,n-k
      do 20 j=i+1,n+1-k
      if(r(i,j).gt.rmax) then
      rmax=r(i,j)
      ii=i
      jj=j
      end if
20    continue
      return
      end

      subroutine cp2(n,k,ii,jj,dmin)
      common /og/x(150,150),d(150,150),nn(150)
      dmin=10000.
      do 20 i=1,n-k
      do 20 j=i+1,n+1-k
      if(d(i,j).lt.dmin) then
      dmin=d(i,j)
      ii=i
      jj=j
      end if
20    continue
      return
      end

      subroutine cd1(n,k,ii,i,nc)
      common /og/x(150,150),r(150,150),nn(150)
      if(i.lt.ii) then
      d1=r(i,ii)
      else
      d1=r(ii,i)
      end if
      d2=r(i,n-k+1)
      d4=abs(d1-d2)
      if(nc.le.2) then
      s1=0.5
      s2=0.5
      if(nc.eq.1) then
      s4=0.5
      else
```

```
s4=-0.5
end if
else if(nc.eq.3) then
s1=(nn(ii)-nn(n+1-k)+0.0)/nn(ii)
s2=1.-s1
s4=0.
end if
ds=s1*d1+s2*d2+s4*d4
if(i.lt.ii) then
r(i,ii)=ds
else
r(ii,i)=ds
end if
return
end

subroutine cd2(n,k,ii,i,nc)
common /og/x(150,150),d(150,150),nn(150)
if(nc.eq.1) then
call cd1(n,k,ii,i,2)
go to 50
else if(nc.eq.2) then
call cd1(n,k,ii,i,1)
go to 50
else if(nc.eq.3) then
call cd1(n,k,ii,i,3)
go to 50
else
d2=d(i,n-k+1)*d(i,n-k+1)
d3=d(ii,n-k+1)*d(ii,n-k+1)
s1=(nn(ii)-nn(n+1-k)+0.0)/nn(ii)
s2=1.-s1
s3=-s1*s2
if(i.lt.ii) then
d1=d(i,ii)*d(i,ii)
d(i,ii)=sqrt(d1*s1+d2*s2+d3*s3)
else
d1=d(ii,i)*d(ii,i)
d(ii,i)=sqrt(d1*s1+d2*s2+d3*s3)
end if
```

```
      end if
50    return
      end

      subroutine flt(n,na)
      common /ofl/r(150),kb(150)
      real k1(150,3),k(150)
      dimension k2(150,2)
      common /fp/xyr(150,4),xyb(150,4)
      common /cs/dz,xs,xf,sa,xmax,xmin
      real ki1,ki11,ki2,ki21,i2,i1,i3
      dz=1.
      xn=0.
      xs=-5.
      xf=5.
      nsf=n/50
      if(nsf.ge.2) then
      xs=xs * 2
      xf=xf * 2
      end if
      xr=xn-dz/2
      if (na.eq.1) then
      xmin=r(2)
      do 10 i=3,n
      xmin=amin1(xmin,r(i))
10    continue
      sa=xf-xs
      do 15 i=1,n
      k1(i,1)=(1.-r(i)) * sa
15    continue
      else
      xmax=r(2)
      xmin=r(2)
      do 20 i=3,n
      xmax=amax1(xmax,r(i))
20    xmin=amin1(xmin,r(i))
      sa=(xf-xs)/xmax
      do 25 i=2,n
25    k1(i,1)=r(i) * sa
      end if
```

```
      k1(1,3)=k1(2,1)
      do 30 i=2,n-1
      if(k1(i,1).gt.k1(i+1,1)) then
      k1(i,3)=k1(i+1,1)
      k2(i,2)=i+1
      else
      k1(i,3)=k1(i,1)
      k2(i,2)=i
      end if
30    continue
      k1(n,3)=k1(n,1)
      k2(n,2)=n
      do 35 i=1,n
35    k(i)=0
      do 80 ii=2,n
      if(ii.eq.2) then
      do 40 i=3,n
      if(k1(i,1).gt.k1(2,1)) then
      k1(2,2)=k1(i,1)
      k2(2,1)=i
      go to 80
      end if
40    continue
      k1(2,2)=k1(2,1)
      k2(2,1)=2
      else if(ii.ne.n) then
      do 45 i=ii-1,2,-1
      if(k1(i,1).ge.k1(ii,1)) then
      ki1=k1(i,1)
      ki11=i
      go to 50
      end if
45    continue
      ki1=1000
      ki11=ii
50    do 55 i=ii+1,n
      if(k1(i,1).gt.k1(ii,1)) then
      ki2=k1(i,1)
      ki21=i
      go to 60
```

```
      end if
55    continue
      ki2=1000
      ki21=ii
60    if(ki1.le.ki2.and.ki11.ne.ii) then
      k1(ii,2)=ki1
      k2(ii,1)=ki11
      else if(ki2.lt.ki1) then
      k1(ii,2)=ki2
      k2(ii,1)=ki21
      else
      k1(ii,2)=k1(ii,1)
      k2(ii,1)=ii
      end if
      else
      do 65 i=n-1,1,-1
      if(k1(i,1).ge.k1(n,1)) then
      k1(n,2)=k1(i,1)
      k2(n,1)=i
      go to 70
      end if
65    continue
      k1(n,2)=k1(n,1)
      k2(n,1)=n
      end if
70    if(k1(ii,2).eq.k1(ii,1).and.k2(ii,1)+1.lt.ii) then
      do 75 i=k2(ii,1)+1,ii-1
      if(k1(i,2).eq.k1(ii,2)) k(i)=1
75    continue
      k(ii)=2
      end if
80    continue
      do 90 l=2,n
      xn=xn+dz
      i1=k1(l,1)
      i2=k1(l,2)
      i3=k1(l,3)
      xyr(l,1)=xr
      xyr(l,2)=i1
      xyr(l,3)=xr
```

```
      xyr(l,4)=i2
      xyb(l,1)=xn
      xyb(l,2)=0.
      xyb(l,3)=xn
      xyb(l,4)=i3
      xr=xr-dz
90    continue
      return
      end

      subroutine plot(n,na)
      common /fp/xyr(150,4),xyb(150,4)
      common /cs/dz,xs,xf,sa,xmax,xmin
      common /ofl/a(150),kb(150)
      dimension x(10),y(11)
      write( * , * ) ' Enter x0,y0,dh,ii :'
      read( * , * ) x0,y0,dh,ii
      dx=0.0
      call in
      call pen(1)
      call fact(10. )
      call setsty('set10.sym')
      xmar=xyr(1,4)
      do 10 i=1,n
      if(xyr(i,4).ge.xmar) xmar=xyr(i,4)
10    continue
      if(n.lt.10) yn=y0+7 * dh+1.
      if(n.ge.10.and.n.lt.100) yn=y0+8 * dh+1.
      if(n.ge.100.and.n.le.1000) yn=y0+9 * dh+1.
      yt=yn+dh
      xmar=xmar+yt
      if(na.eq.1) then
      do 20 i=1,10
      y(i)=yt+dx * sa
      x(i)=1.-dx
20    dx=dx+0.25
      y(i)=y(i-1)+abs(xmin) * sa
      else
      do 30 i=1,10
      y(i)=yt+dx * xmax * sa
```

```
      x(i)=dx * xmax
30    dx=dx+0.25
      y(i)=y(i-1) + abs(xmin) * sa
      end if
      do 40 i=1,10
      if(y(i).lt.xmar) then
      call movea(x0+0.75 * dz,y(i))
      call linea(x0+dz,y(i))
      call numb1(x0+0.5 * dz,y(i)-3 * dh,dh,90.,x(i))
      end if
40    continue
      xt=x0+dz
      call movea(xt,y(1))
      if(y(11).lt.xmar) then
      call linea(xt,y(11))
      else
      call linea(xt,xmar)
      end if
      xt=xt+dh
      call movea(xt,yt)
      call linea(xt,yt+xyb(2,4))
      do 50 i=1,n
      if(yt+xyb(i,4).eq.yt+xyb(2,4))
     # call linea(xt+xyb(i,3),yt+xyb(2,4))
50    continue
      xn=xt+0.25 * dz
      xr=xt-0.25 * dz
      do 80 i=1,n
      if(ii.eq.0) then
      if(i.ge.2) call numb1(xr,y0+1.,dh,90.,a(i))
      xr=xr+dz
      end if
      if(kb(i).lt.10) then
      call numb2(xn,yn,dh,90.,kb(i))
      else
      if(kb(i).ge.10.and.kb(i).lt.100) then
      call numb2(xn,yn-dh,dh,90.,kb(i))
      else
      call numb2(xn,yn-2 * dh,dh,90.,kb(i))
      end if
```

```
      end if
      xn=xn+dz
      if(xyb(i,2).ne.xyb(i,4)) then
      call movea(xt+xyb(i,1),yt+xyb(i,2))
      call linea(xt+xyb(i,3),yt+xyb(i,4))
      end if
      if(xyr(i,2).ne.xyr(i,4)) then
      call movea(xt-xyr(i,1),yt+xyr(i,2))
      call linea(xt-xyr(i,3),yt+xyr(i,4))
      end if
      do 70 j=i+1,n
      if(xyr(i,4).eq.xyr(j,4)) then
      call movea(xt-xyr(i,3),yt+xyr(i,4))
      call linea(xt-xyr(j,3),yt+xyr(j,4))
      end if
      if(xyb(i,4).eq.xyb(j,4)) then
      call movea(xt+xyb(i,3),yt+xyb(i,4))
      call linea(xt+xyb(j,3),yt+xyb(j,4))
      end if
      if(xyb(j,4).eq.xyr(i,4)) then
      call movea(xt-xyr(i,3),yt+xyr(i,4))
      call linea(xt+xyb(j,3),yt+xyb(j,4))
      end if
      if(xyr(j,4).eq.xyb(i,4)) then
      call movea(xt+xyb(i,3),yt+xyb(i,4))
      call linea(xt-xyr(j,3),yt+xyr(j,4))
      end if
70    continue
80    continue
      return
      end
```

二、最优分割程序

本程序以段内离差平方和为分类统计量，对具有 m 个变量的 n 个有序样品进行最优 2-k 段分割，并绘制 S_n-k 图。对程序的主要参数、符号及其使用方法说明如下：

（一）主要参数及符号

1. 参数

k——最大分割段数；

n1——参加分割的起始样品号；

n2——参加分割的终止样品号；

mk——参加分割的变量数。

2. 符号

x——参加分割的数据矩阵；

mv　　存放参加分割的变量号矩阵；

nkp——存放分割点的矩阵；

d——段内变差矩阵；

w——段内离差平方和矩阵。

4. 子程序

cd1——计算段内变差和段内离差平方和子程序；

norm——数据极差正规化子程序；

plot1——绘制 S_n k 图子程序；

plot2——绘制变量曲线子程序。

（二）操作说明

1. 数据文件

数据文件由 $n\times m$ 个观测值构成，格式与聚合聚类中的文件格式相同。

2. 操作说明

在 DOS 操作系统下，键入最优分割目标程序名 zyfg，回车后的具体操作步骤如下：

(1) 首先键入绘图坐标原点 $x0$、$y0$、及样品总数 n 和变量数 m(Input $x0$,$y0$,n,m)。

(2) 输入数据文件名(Input your adta file name)。

(3) 键入最大分割段数 k(Input maxmum number of group divided)。

(4) 键入参加分割的起、止样品号 $n1$ 和 $n2$。(Input n1,n2)。

(5) 键入参加分割的变量数 mk(Input mk)。

(6) 键入分割所采用变量的序号(Input variable order)。

(7) 键入绘图文件名(Enter plot filename)。

(8) 最后，k 段分割结束后，先后显示：

To change the variable number (y/n)?

To change the sample number (y/n)?

To change maxmum number of group divided (y/n)?

若键入 y，将分别返回变量数、样品数和最大分段数 k 选择处；若键入 n，则继续往下执行，直至结束。

3. 主要输出结

在文件 kfg.dat 中记录了 2-k 分割的分割点、分段结果(段内的样品序号)、段内离差平方和等。另外、在用户给定的绘图文件中中存贮了 S_n-k 图及有序样品的分段图。

（三）源程序

1. 最优分割程序框图

本程序框图如图 4-5 所示。

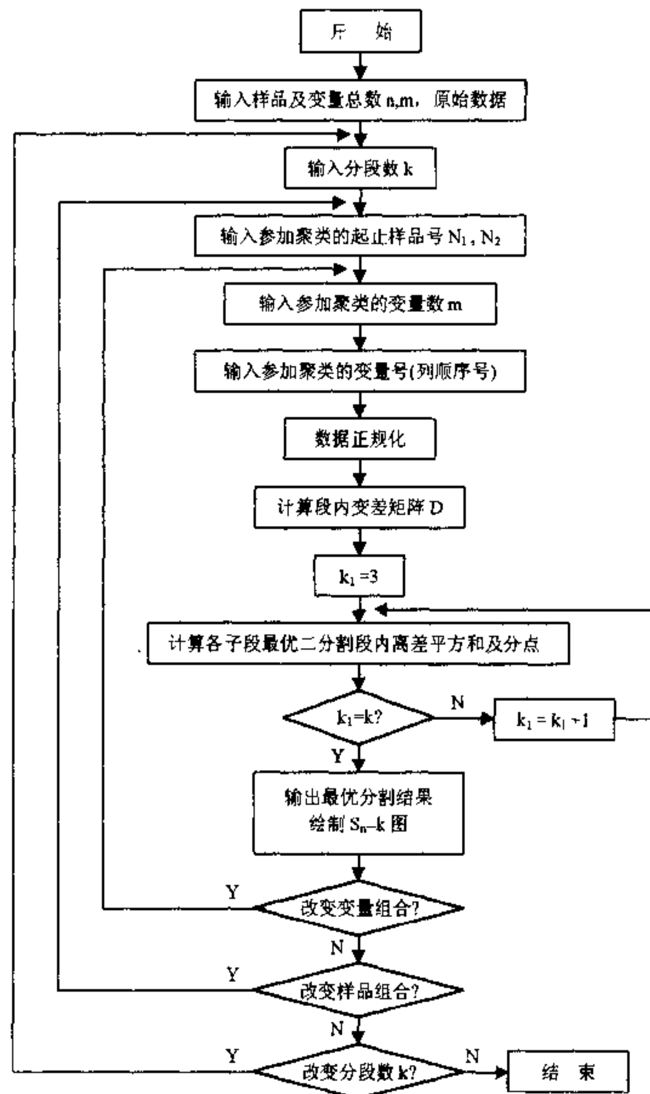


图 4-5 最优分割程序框图

2. 最优分割源程序

```
c          program zyfg.for
$debug
           dimension x(100,6),xm(100,6),d(100,100),kf(100)
           common /on/mv(6),nk(100),nkp(100),nw(100,100)
           common /oc/w(100,100),z(100,6).x0,y0,dy
           common/yxk/yk(100),kx(100)
           equivalence (d(1,1),w(1,1))
           character fname * 10,fam * 10,yn
           fam='kfg.dat'
           write( * , * ) ' Input x0,y0,dy,n,m '
```

```
      read( * , * ) x0,y0,dy,n,m
      write( * ,'(a)') ' Input your data file name '
      read( * ,'(a)') fname
      open(2,file=fam)
      open(3,file=fname)
      do 100 i=1,n+10
100   read(3, * ,end=120,err=110) (xm(i,j),j=1,m)
110   write( * ,'(a,i3)') ' Err of file:i=',i
      stop
120   write( * ,'(a,i3,a,i2)')' End of file:i=',i-1,' j=',j-1
130   write( * ,'(a)') ' Input maxmum number of group divided '
      write( * ,'(a)') ' k=? '
      read( * , * ) k
140   write( * ,'(a)') ' Input n1,n2: n1=? ,n2=? '
      read( * , * ) n1,n2
      n=n2-n1+1
150   write( * ,'(a)') ' Input mk: mk=? '
      read( * , * ) mk
      write( * ,'(a)') ' input variable order '
      read( * , * ) (mv(i),i=1,mk)
      ii=0
      do 170 i=n1,n2
      ii=ii+1
      do 170 j=1,mk
170   x(ii,j)=xm(i,mv(j))
      do 180 i=1,n
      do 180 j=1,mk
180   z(i,j)=x(i,j)
      call norm(n,mk)
      call cd1(n,mk)
      do 190 k1=2,k
190   nw(k1,k1)=k1-1
      do 210 m1=3,n
      c=100000.
      do 200 j=1,m1-1
      dc=d(1,j)+d(j+1,m1)
      if(dc.lt.c) then
      c=dc
      nc=j
      end if
```

```
200        continue
           w(m1,2)=c
210        nw(2,m1)=nc
           do 240 k1=3,k
           if(k1.eq.k) then
           k2=n
           else
           k2=k1+1
           end if
           do 230 m1=k2,n
           c=100000.
           do 220 j=k1-1,m1-1
           dc=w(j,k1-1)+d(j+1,m1)
           if(dc.lt.c) then
           c=dc
           nc=j
           end if
220        continue
           w(m1,k1)=c
230        nw(k1,m1)=nc
240        continue
c          write(*,'(a)') ' Matrix W '
c          do 250 i=2,n
c          if(i.le.k) then
c          i1=i
c          else
c          i1=k
c          end if
c250       write(*,'(10(2x,6f10.4/))') (w(i,j),j=2,i1)
           do 300 k1=2,k
           nk(2)=nw(k1,n)
           do 260 j=3,k1
260        nk(j)=nw(k1+2-j,nk(j-1))
           do 270 i=2,k1
270        nkp(i)=nk(k1-i+2)+n1-1
           write(2,'(a,i2,a,f10.4)')'k=',k1,' Class Diameter sum=',w(n,k1)
           write(2,'(a,100(10i4/))')'Dividing points:',(nkp(i),i=2,k1)
           nk(1)=n
           nk(k1+1)=0
           write(2,'(a)') ' Samples of each class '
```

```
      do 300 lk=1,k1
      l1=nk(k1-lk+2)+1
      l2=nk(k1-lk+1)
      do 290 i=l1,l2
290   kf(i)=i+n1-1
      jk=l2-l1+1
      md=int(jk/10.)
      if(md.eq.0) then
      if(jk.eq.1) write(2,'(a,i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.2) write(2,'(a,2i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.3) write(2,'(a,3i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.4) write(2,'(a,4i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.5) write(2,'(a,5i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.6) write(2,'(a,6i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.7) write(2,'(a,7i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.8) write(2,'(a,8i3,a)') '{',(kf(i),i=l1,l2),'}'
      if(jk.eq.9) write(2,'(a,9i3,a)') '{',(kf(i),i=l1,l2),'}'
      go to 300
      end if
      kj=l1+10*md-1
      kkj=l2-kj
      write(2,'(a,5(10i3/2x))') '{',(kf(i),i=l1,kj)
      if(kkj.eq.1) write(2,'(i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.2) write(2,'(2i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.3) write(2,'(3i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.4) write(2,'(4i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.5) write(2,'(5i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.6) write(2,'(6i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.7) write(2,'(7i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.8) write(2,'(8i3,a)') (kf(i),i=kj+1,l2),'}'
      if(kkj.eq.9) write(2,'(9i3,a)') (kf(i),i=kj+1,l2),'}'
300   continue
      do 310 i=2,k
      kx(i)=i
310   yk(i)=w(n,i)
      call in
      call fact(10.)
      call plot1(x0,y0,k,5.)
      call in
      call plot2(n,m,k)
```

```
      write( * ,'(a)') ' To change the variable number (y/n) ? '
      read( * ,'(a)') yn
      if(yn.eq.'y'.or.yn.eq.'Y') go to 150
      write( * ,'(a)') ' To change the sample number (y/n) ? '
      read( * ,'(a)') yn
      if(yn.eq.'y'.or.yn.eq.'Y') go to 140
      write( * , * ) ' To change maxmun number of group divided (y/n) ? '
      read( * ,'(a)') yn
      if(yn.eq.'y'.or.yn.eq.'Y') go to 130
      stop
      end

      subroutine norm(n,m)
      common/oc/w(100,100),x(100,6),x0,y0,dy
      do 120 j=1,m
      xmax=x(1,j)
      xmin=x(1,j)
      do 100 i=2,n
      if(x(i,j).gt.xmax) xmax=x(i,j)
      if(x(i,j).lt.xmin) xmin=x(i,j)
100   continue
      d=xmax-xmin
      do 110 i=1,n
110   x(i,j)=(x(i,j)-xmin)/d
120   continue
      return
      end

      subroutine cd1(n,m)
      dimension d(100,100)
      common /oc/w(100,100),x(100,6),x0,y0,dy
      equivalence (d(1,1),w(1,1))
      do 100 i=1,n
100   d(i,i)=0.
      do 120 j=2,n
      sum=0.
      do 120 i=j-1,1,-1
      c=0.
      do 110 l=1,m
110   c=c+(x(i,l)-x(j,l))*(x(i,l)-x(j,l))
```

```
      sum=sum+c
120   d(i,j)=d(i,j-1)+sum
      do 130 i=1,n-1
      do 130 j=i+1,n
130   d(i,j)=d(i,j)/(j-i+1)
      return
      end

      subroutine plot1(x0,y0,k,r)
      common/yxk/yk(100),kx(100)
      dh=0.4
      call text(x0+k-1.4,y0-0.8,dh,0.,'K')
      call text(x0+0.1,y0+yk(2)/r+0.2,dh,0.,'Sn')
      call movea(x0,y0)
      call linea(x0+k-1,y0)
      call linea(x0+k-1.4,y0-0.1)
      call movea(x0+k-1,y0)
      call linea(x0+k-1.4,y0+0.1)
      call movea(x0,y0)
      call linea(x0,y0+yk(2)/r+1)
      call linea(x0-0.1,y0+yk(2)/r+0.6)
      call movea(x0,y0+yk(2)/r+1.)
      call linea(x0+0.1,y0+yk(2)/r+0.6)
      yy=y0-2*dh
      xx=x0-1
      do 100 i=2,k
      xx=xx+1
100   call numb2(xx,yy,dh,0.,i)
      call movea(x0,y0+yk(2)/5)
      xx=x0-1.
      do 200 i=2,k
      xx=xx+1.
      yy=y0+yk(i)/5
200   call linea(xx,yy)
      return
      end

      subroutine plot2(n,m,k)
      common/on/mv(6),nk(100),nkp(100),nw(100,100)
      common/oc/w(100,100),z(100,6),x0,y0,dy
```

```
      y0=y0-2
      dx=1.5
      dx1=1.2
      do 200 i=1,m
      x=x0+(i-1)*dx
      y=y0
      call movea(x,y)
      call linea(x+dx1,y)
      call movea(x,y)
      call linea(x,y-(n-1)*dy)
      call movea(x+z(1,i),y)
      do 100 j=2,n
      y=y-dy
100   call linea(x+z(j,i),y)
200   continue
      x=x0+dx*m
      do 300 i=2,k
      y=y0-dy*nkp(i)
      call movea(x0,y)
300   call linea(x,y)
      return
      end
```

§5 应用算例

【例1】 在油气地表化探工作中，一般都要对每个样品分析甲烷、乙烷、丙烷、紫外、荧光、总烃和重烃等20余项指示地下油气的指标。这些指标间既有相对的独立性、又存在着一定的成因联系，因此，可借助聚类分析对它们进行分类，从中选择出有代表性的指标。

中国东部某地区1 600余个样品29项化探指标的聚类谱系图如图4-6所示。在谱系图上，取相关系数$r=0.75$，可把29项指标分为13类。其中成因联系密切的有酸解烃类、紫外和荧光类。

酸解烃类包含了全部的酸解烃指标。其中甲烷与总烃以$r=0.98$聚为一类，其原因在于甲烷占了总烃的绝大部分。甲烷在油气藏中浓度最高，运移能力强，到达地表的数量最多。重烃的综合性强，易被土壤吸附，直接反映烃场浓度的大小和地下油气的性质。因此，可把它们作为油气化探的重要指标。

紫外和荧光类包括了除紫外216外的全部紫外和荧光指标。它们都是检测油气藏中芳烃混合物的直接指标，因此，可从紫外和荧光中各选一个作为代表性指标。

【例2】 应用聚类分析方法可对不同勘探程度的含油气盆地进行远景评价，其具体做法是：

选取一定数量的含油气盆地(其中包括勘探程度较高的和待评价的盆地)及盆地间可类比

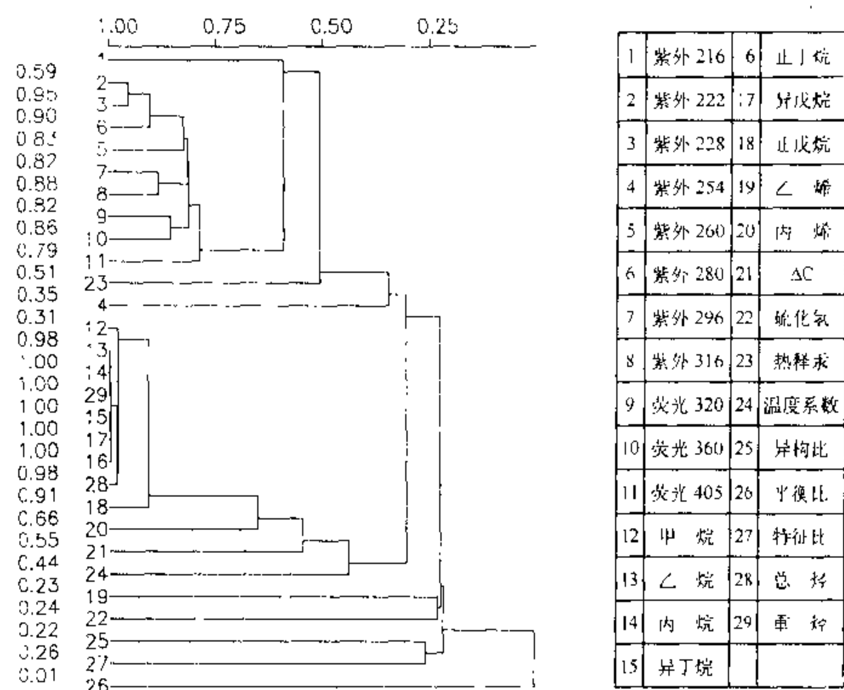


图 4-6 中国东部某地区油气化探指标聚类谱系图及指标

的地质参数(数值型和非数值型)作聚类分析,根据聚类结果,由已知含油气远景的盆地推测待评价盆地的含油气远景。

本例中选取了世界上100个含油气盆,可类比的地质参数有5类共30项,原始数据见表4-2。对各含油气盆地中的非数值(如盆地类型)型参数,在计算前必须作数值化处理,处理方法是:

① 盆地类型:分内陆、沿海、海湾和海洋,按是为1,不是为0编码;

② 地质时代:分第三纪、白垩纪、侏罗纪、三迭纪、二迭纪、石炭纪、泥盆纪、志留纪、奥陶纪和寒武纪共10项,按有为1,无为0编码;

③ 岩性:分砂岩、碳酸盐岩、火成岩和基岩四项,按有为1,无为0编码;

在表4-2中,地质参数按列的排列顺序是盆地编号、面积(10万平方公里计)、沉积时代(10个时代)、沉积厚度(万米计)、储层地质时代(10个时代)、储层岩性(4种岩性)、盆地类型(4种类型)共31项。

对原始数据作标准差标准化后,分别采用相关系数近邻联接法和欧氏距离最短距离法进行系统聚类,结果如图4-7a和b所示。

表 4-2 含油气盆地地质类比参数

序号	面积	沉积时代	厚度	储层时代	储层岩性	盆地类型
1	0.22	1100000000	0.40	1100000000	1100	0100
2	0.40	1100000000	1.60	0100000000	1000	0100
3	1.00	1100000000	0.80	1000000000	1000	0100
4	2.20	1111111111	0.80	0100000000	1000	0100
5	1.00	0000011111	0.10	0000000101	1100	1000

续表 4-2

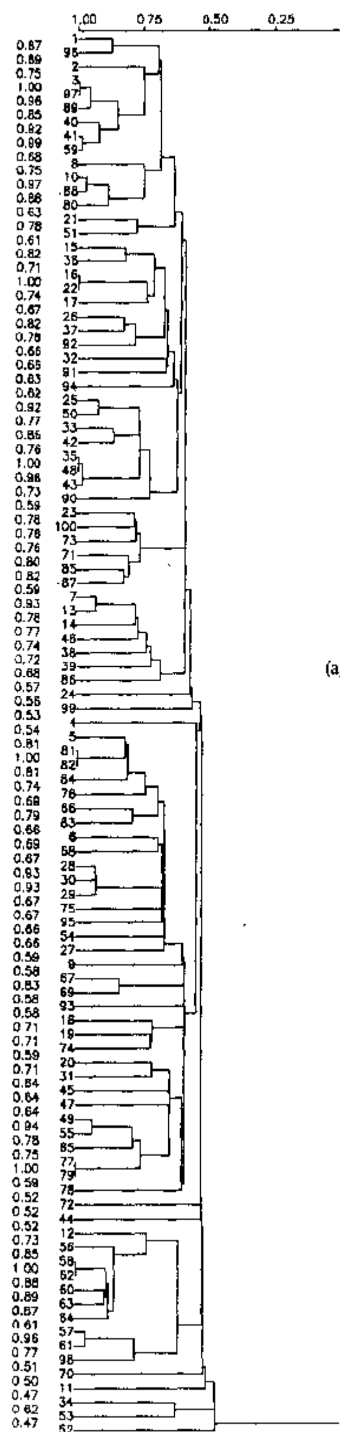
序号	面积	沉积时代	厚度	储层时代	储层岩性	盆地类型
6	1.10	0000011111	0.40	0000011000	1000	1000
7	0.07	1111000000	0.30	0010000000	0100	0100
8	0.19	1111000000	0.30	1111000000	1101	0100
9	4.50	0111111111	0.70	0001000011	1000	0100
10	4.00	1100000000	0.50	1100000000	1101	0100
11	4.00	1111111000	0.40	1111111000	1100	0010
12	0.20	1111000000	0.60	1001000000	1100	0001
13	1.20	1111000000	0.50	0110000000	0100	0100
14	0.90	1111000000	1.00	0110000000	0100	1000
15	1.30	1111000000	0.33	0110000000	1100	1000
16	0.58	1111000000	0.45	1111000000	1100	1000
17	0.10	1111100000	0.30	1011000000	1100	1000
18	0.40	0111111000	0.30	0011010000	1100	0001
19	5.00	1111111000	0.60	1111100000	1100	0001
20	0.80	1111110000	0.80	0111100000	1100	1000
21	0.70	1000000000	0.45	1000000000	1001	1000
22	2.00	1111000000	0.50	1111000000	1100	1000
23	3.40	1111000000	1.00	1111000000	1100	0010
24	1.10	1111000000	1.20	1110000000	1100	0100
25	0.26	1000000000	0.40	1000000000	1000	1000
26	1.10	1111000000	0.80	1100000000	1000	1000
27	0.40	1111111111	0.40	0000000111	1100	0010
28	1.35	0111111000	0.60	0000111000	1100	1000
29	6.90	0001111000	0.60	0000111000	1100	1000
30	2.00	0011111000	0.35	0000111000	1100	1000
31	5.00	1111111000	1.50	0111100000	1000	1000
32	5.20	1101000000	1.00	1110000000	1100	1000
33	2.60	1010000000	1.20	1000000000	1000	1000
34	32.80	1101111111	1.10	1110000000	1100	0010
35	1.20	1110000000	0.80	1000000000	1000	1000
36	5.30	1110000000	0.50	0110000000	1100	1000
37	0.40	1111000000	0.80	1100000000	1100	1000
38	1.30	1110000000	0.70	1100000000	0100	1000
39	3.60	1101000000	0.60	1000000000	0100	1000
40	0.40	1000000000	0.50	1000000000	1000	0100
41	2.50	1000000000	1.00	1000000000	1000	0100
42	0.50	1010000000	0.60	1000000000	1000	1000

续表 4-2

序号	面积	沉积时代	厚度	储层时代	储层岩性	盆地类型
43	1.20	1110000000	1.00	1000000000	1000	1000
44	6.10	1111111000	1.00	1010000000	1000	1000
45	1.60	1111110000	1.30	1001100000	1000	1000
46	0.50	1110000000	0.75	0010000000	0000	1000
47	2.90	0101111000	0.70	0011000000	1000	1000
48	0.30	1110000000	0.80	1000000000	1000	1000
49	2.10	0101111111	1.00	0011100000	1100	1000
50	0.80	1000000000	0.80	1000000000	1000	1000
51	3.80	1000000000	0.50	1000000000	1001	0010
52	12.70	0110000000	1.80	0100000000	1000	1000
53	33.00	1101000000	0.80	0110000000	1000	1000
54	5.50	0010000111	1.20	0000000001	1100	1000
55	4.60	0101111111	1.00	0011100000	1000	1000
56	0.50	1100000000	0.80	1000000000	1000	0001
57	0.80	1000000000	0.60	1000000000	1010	0001
58	0.16	1000000000	0.50	1000000000	1000	0001
59	1.20	1000000000	1.20	1000000000	1000	0100
60	2.30	1000000000	0.75	1000000000	1100	0001
61	0.40	1000000000	0.60	1000000000	1110	0001
62	0.35	1000000000	0.60	1000000000	1000	0001
63	1.50	1010000000	0.90	1000000000	1100	0001
64	3.50	1110000000	1.00	1000000000	1100	0001
65	1.04	1101111110	1.50	0011100000	1000	1000
66	1.55	0000001111	0.90	0000000010	1000	1000
67	1.27	0001100000	0.37	0000100000	1000	1000
68	0.28	0000011000	0.50	0000001000	1000	1000
69	1.94	0001100000	0.90	0001100000	1000	1000
70	2.70	0110000000	0.24	0010000000	1000	1000
71	0.40	1100000000	0.70	1100000000	1000	0010
72	3.30	1111110000	0.90	1111010000	1100	0100
73	0.39	1111000000	0.90	1000000000	1000	0010
74	3.10	1111110000	0.45	0111110000	1100	0001
75	6.00	0111111111	0.60	0101111000	1100	1000
76	6.50	1111111111	0.46	0011011110	1100	1000
77	0.60	1111111111	0.50	0111100000	1000	1000
78	6.00	1111111111	1.80	1111110000	1100	1000
79	1.30	1111111111	0.50	0111100000	1000	1000

续表 4-2

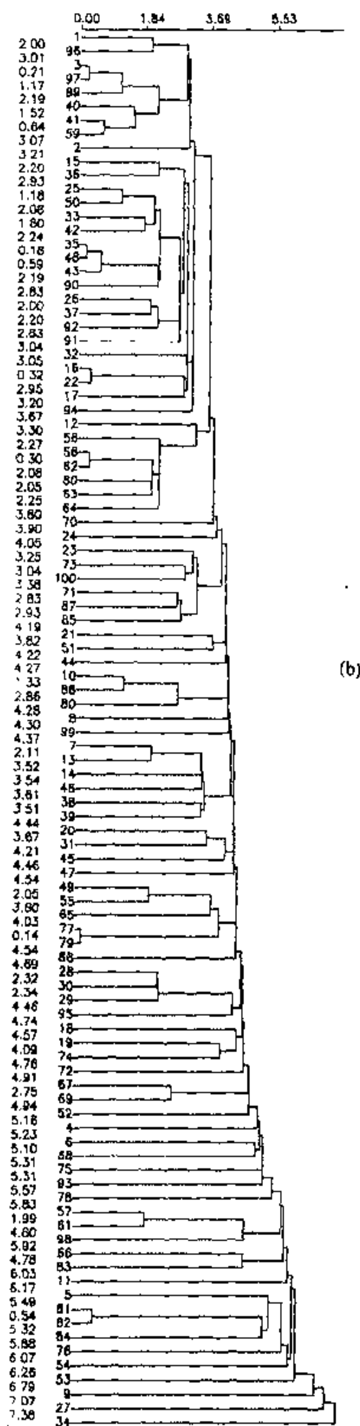
序号	面积	沉积时代	厚度	储层时代	储层岩性	盆地类型
80	1.80	1100000000	1.60	1100000000	1001	0100
81	3.10	0000011111	0.45	0000011111	1100	1000
82	4.70	0000011111	0.60	0000011111	1100	1000
83	11.30	0000111111	1.00	0000100010	1100	1000
84	3.20	0000111111	0.60	0000111110	1100	1000
85	11.00	1110000000	1.00	1110000000	1100	0010
86	0.62	1110000000	0.30	0110000000	0100	0010
87	1.60	1110000000	0.70	1100000000	1100	0010
88	0.85	1100000000	0.90	1100000000	1101	0100
89	1.04	1100000000	1.20	1000000000	1000	0100
90	2.00	1100000000	0.50	1000000000	1000	1000
91	1.20	1100000000	0.80	1100000000	1100	1000
92	2.60	1111000000	0.50	0100000000	1000	1000
93	12.00	1111111111	0.40	0000001000	1000	1000
94	1.80	1111110000	0.90	0100000000	1000	1000
95	2.80	1111111000	0.60	0100011000	1000	1000
96	0.50	1100000000	0.45	0100000000	1100	0100
97	0.25	1100000000	0.75	1000000000	1000	0100
98	1.30	1001000000	0.35	1001000000	1010	1000
99	1.40	0111000000	0.40	0110000000	1000	0100
100	2.10	1111000000	0.60	1110000000	1000	0010



(a)

1		51	渤海湾
2		52	松辽盆地
3	尼日利亚沿海	53	西西伯利亚
4		54	
5		55	
6	波利尼亚克	56	萨哈林
7		57	
8		58	
9	三叠盆地	59	
10	锡尔特盆地	60	苏门答腊
11	苏伊士-红海	61	北爪哇盆地
12		62	沙捞越
13		63	东加里曼丹
14		64	
15		65	佩思盆地
16		66	
17		67	库珀盆地
18		68	
19	北海	69	
20	荷兰-西德盆地	70	
21	维也那盆地	71	吉普斯兰
22		72	
23		73	库克湾
24		74	
25		75	阿尔伯达盆地
26	前喀尔巴阡山	76	
27		77	粉河盆地
28		78	落基山诸盆地
29	伏尔加-乌拉山	79	丹佛盆地
30	伯朝拉盆地	80	加里福尼亚诸盆地
31		81	
32		82	阿巴拉契亚
33	南里海	83	北美地台
34	波斯湾	84	西德克萨斯(二)
35		85	墨西哥湾
36	布哈提	86	
37		87	维拉克鲁斯-塔巴斯克
38		88	马拉开波-法尔康
39		89	
40		90	
41		91	
42		92	普图马约
43	柴达木盆地	93	
44	塔里木盆地	94	
45	准噶尔盆地	95	圣克鲁斯
46	吐鲁番盆地	96	
47	陕甘宁盆地	97	瓜亚及尔
48	九泉盆地	98	
49	四川盆地	99	
50	江汉盆地	100	

(a) 近邻联接法聚类谱系图



(b)

1		51	渤海湾
2		52	松辽盆地
3	尼日利亚沿海	53	西西伯利亚
4		54	
5		55	
6	波利尼亚克	56	萨哈林
7		57	
8		58	
9	三叠盆地	59	
10	锡尔特盆地	60	苏门答腊
11	苏伊士-红海	61	北爪哇盆地
12		62	沙捞越
13		63	东加里曼丹
14		64	
15		65	佩思盆地
16		66	
17		67	库珀盆地
18		68	
19	北海	69	
20	荷兰-西德盆地	70	
21	维也那盆地	71	吉普斯兰
22		72	
23		73	库克湾
24		74	
25		75	阿尔伯达盆地
26	前喀尔巴阡山	76	
27		77	粉河盆地
28		78	落基山诸盆地
29	伏尔加-乌拉山	79	丹佛盆地
30	伯朝拉盆地	80	加里福尼亚诸盆地
31		81	
32		82	阿巴拉契亚
33	南里海	83	北美地台
34	波斯湾	84	西德克萨斯(二)
35		85	墨西哥湾
36	布哈提	86	
37		87	维拉克鲁斯-塔巴斯克
38		88	马拉开波-法尔康
39		89	
40		90	
41		91	
42		92	普图马约
43	柴达木盆地	93	
44	塔里木盆地	94	
45	准噶尔盆地	95	圣克鲁斯
46	吐鲁番盆地	96	
47	陕甘宁盆地	97	瓜亚及尔
48	九泉盆地	98	
49	四川盆地	99	
50	江汉盆地	100	

(b) 最短距离法系统聚类谱系图

图 4-7　100 个含油气盆地聚类谱系图及盆地名称

由于类比参数的局限性，所以说这个算例仅是方法上的演习，结论仅供参考。尽管如此，所得结论仍能说明一定的问题。

【例 3】 四川盆地中某探井 2670～2720m 井段有自然伽玛、中子、密度、声波时差、浅侧向电阻率、深侧向电阻率 6 种测井资料。在该井段内等间距地取了 56 个样品，构成一个有序样品段。6 种测井资料的采样值见表 4-3。

对表 4-3 中的数据进行了最优 5 分割。分割点为 13、17、26、47，各子段内样品号为：

(1-13)、(14-17)、(18-26)、(27-47)、(48-56)。

从图 4-8 上可以看出，5 次分割可将大段的岩性分开。在图 4-9 上，5 次分割对应的段内离差平方和已明显变小，曲线变缓。这就是说，在曲线变缓时对应的分段数 k 可作为大段岩性划分的段数。

表 4-3 六种测井资料的原始数据(据赵旭东)

样品号	自然伽玛	中 子	密 度	声波时差	浅侧向	深测向
1	28.000	3.000	2.840	49.000	700.000	700.000
2	60.000	7.500	2.780	53.000	90.000	65.000
3	10.000	0.750	2.920	50.000	1100.000	1000.000
4	48.000	0.000	2.830	52.000	110.000	88.000
5	10.000	1.000	2.860	50.000	1600.000	1900.000
6	35.000	9.750	2.740	48.000	100.000	150.000
7	16.000	9.000	2.700	48.000	140.000	155.000
8	25.000	12.000	2.670	49.000	28.000	28.000
9	52.000	3.000	2.680	51.000	110.000	100.000
10	24.000	0.750	2.720	52.000	200.000	200.000
11	20.000	0.000	2.710	48.000	1300.000	1300.000
12	25.000	1.500	2.750	49.000	1000.000	1000.000
13	45.000	4.500	2.830	48.500	220.000	240.000
14	3.000	−1.500	2.960	51.500	1800.000	1800.000
15	26.000	−1.600	2.890	52.000	1600.000	1650.000
16	45.000	−1.400	2.920	52.000	1900.000	1850.000
17	4.000	3.000	2.800	50.000	1000.000	1500.000
18	5.000	10.500	2.580	56.000	95.000	160.000
19	10.000	10.000	2.600	57.000	54.000	80.000
20	19.000	2.200	2.680	50.000	150.000	300.000
21	15.000	1.600	2.660	52.500	80.000	135.000
22	20.000	7.000	2.570	56.000	55.000	80.000
23	19.000	7.600	2.650	55.000	40.000	95.000
24	15.000	11.000	2.650	61.000	25.000	33.000
25	10.000	5.500	2.670	53.500	78.000	100.000
26	16.000	4.500	2.650	52.500	40.000	54.000
27	14.000	0.750	2.660	48.500	1000.000	800.000
28	17.000	1.500	2.650	48.500	1600.000	1100.000
29	15.000	2.300	2.670	48.500	600.000	650.000
30	25.000	6.700	2.700	48.000	160.000	210.000
31	16.000	1.500	2.680	47.500	100.000	600.000
32	12.000	3.000	2.690	47.500	300.000	550.000
33	17.000	4.500	2.700	48.000	210.000	300.000
34	11.000	1.500	2.710	42.000	400.000	560.000
35	13.000	0.000	2.680	42.000	300.000	500.000
36	20.000	0.750	2.680	42.000	550.000	670.000

续表 4-3

样品号	自然伽玛	中子	密度	声波时差	浅侧向	深侧向
37	20.000	0.800	2.700	42.000	400.000	560.000
38	15.000	3.800	2.750	41.500	570.000	680.000
39	14.000	4.500	2.825	41.000	350.000	430.000
40	13.000	4.300	2.774	41.000	400.000	440.000
41	24.000	4.500	2.760	41.000	650.000	700.000
42	21.000	−0.800	2.960	50.000	1900.000	1600.000
43	25.000	1.500	2.890	54.000	900.000	700.000
44	10.000	0.750	2.900	50.000	1250.000	830.000
45	10.000	0.000	2.900	52.500	1000.000	750.000
46	15.000	8.800	2.750	55.000	50.000	32.000
47	79.000	0.300	2.935	51.500	1500.000	960.000
48	9.000	11.800	2.790	60.000	23.000	23.000
49	15.000	10.500	2.775	60.000	30.000	90.000
50	15.000	1.300	2.840	51.000	115.000	30.000
51	64.000	17.800	2.640	58.000	30.000	25.000
52	62.000	11.000	2.700	57.000	51.000	28.000
53	94.000	18.000	2.725	60.000	26.000	18.000
54	53.000	13.500	2.800	57.500	34.000	21.000
55	61.000	12.000	2.775	60.000	29.000	17.000
56	10.000	9.000	2.835	57.500	31.000	24.000

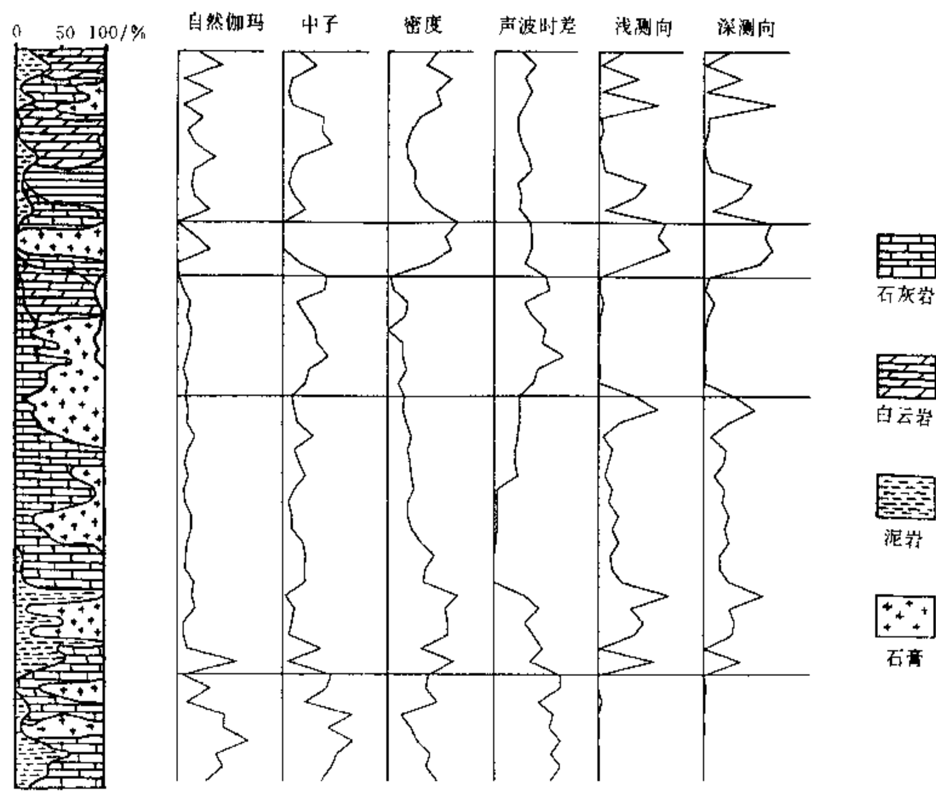


图 4-8 测井解释剖面与 5 段分割结果对比图

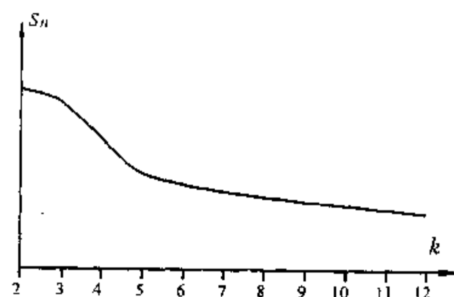


图 4-9　2-12 段分割的 S_n-k 图

习　题

1. 什么是聚类(点群)分析？阐述其基本思想。
2. 在聚类分析中，为何要对原始数据作数据变换？常用的数据变换方法有哪几种？试写出变换公式。
3. 写出常用的几种聚类统计量，并说明如何用它们来刻划样品或变量之间的亲疏关系？
4. 什么是最优分割法？对有序数列最优分割的统计量是什么？
5. 试述最优分割的具体计算步骤。
6. 利用聚类分析程序对例 2 中的数据进行聚类分析。
7. 试用最优分割程序对例 3 中的数据进行最优 5 分割。

第五章　判别分析

在地质研究工作中，经常碰到样品类型归属问题。例如，钻穿的储层是油层、气层、干层还是水层；某块岩石样品是河流相还是属于湖泊相的沉积物；探区中地质圈闭的含油气地质条件是有利的，还是不利的等。上述问题的共同特点就是确定一个样品是属于已知类型中的哪一类，即对样品进行归类。

为叙述方便，在此把样品所属的类统称为总体。判别分析就是根据从已知的 G 个总体中所取出的 G 组样品的观测值，建立样品总体与样品变量之间的定量关系，即判别函数的一种多元统计分析方法。当 $G=2$ 时，称为两总体判别分析，当 $G>2$ 时，称为多总体判别分析。从算法上讲，判别分析通常又有多总体判别分析和逐步判别分析之分。

§1　两总体判别分析

两总体判别分析是根据总体 A、B 的两组样品观测值，建立用于判定样品 X（X 属于 A 或者 B）所属总体的线性判别函数的多元统计分析方法。

一、线性判别函数

设从总体 A 和 B 中分别取出了 n_a 和 n_b 个样品，并且每个样品有两个变量 x_1 和 x_2。在 x_1、x_2 坐标系中作 n_a 和 n_b 个样品的散点图，如图 5-1 所示。对于来自 A 或者 B 的一个新样品 X 来说，如果采用变量 x_1、x_2 确定它的总体，由图 5-1 可见，当 x_1 和 x_2 的值分别落在区间 (a,b) 和 (c,d) 内时，样品 X 的总体是难以判定的。如果把坐标系旋转，得 y、z 新坐标系。新的变量 y 可以把 A、B 两个总体分开。新变量 y 的形式为

$$y=c_1x_1+c_2x_2$$

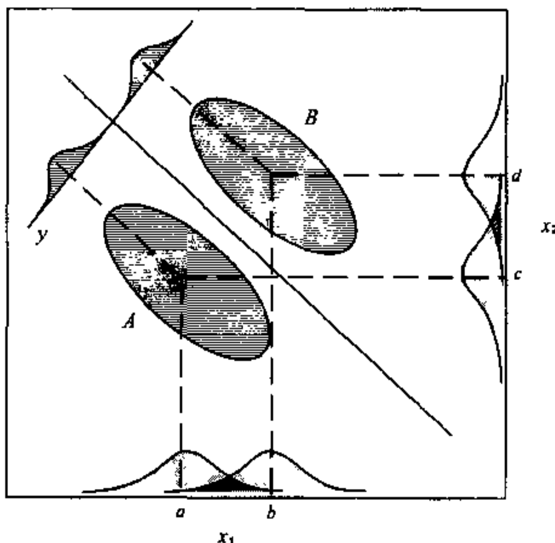


图 5-1　判别分析示意图

一般情况下，若样品有 m 个变量，那么 y 的形式为

$$y=c_1x_1+c_2x_2+\cdots+c_mx_m \quad (5\text{-}1)$$

y 是由 $x_i(i=1,2,\cdots,m)$ 线性组合而成的一个综合性指标，它是 $m+1$ 维空间中的一个平面，称式(5-1)为线性判别函数，其中 $c_1,c_2,\cdots,c_m$ 叫做判别系数。

二、确定判别系数

（一）原始数据

设从总体 A 和 B 中分别取了 n_a 和 n_b 个样品，每个样品有 m 个变量，它们的观测值分别记为 $x_{ij}(a)$ 和 $x_{kj}(b)$ $(i=1,2,\cdots,n_a;k=1,2,\cdots,n_b)$，即

$$\begin{matrix} x_{11}(a) & x_{12}(a) & \cdots & x_{1m}(a) & \quad & x_{11}(b) & x_{12}(b) & \cdots & x_{1m}(b) \\ x_{21}(a) & x_{22}(a) & \cdots & x_{2m}(a) & & x_{21}(b) & x_{22}(b) & \cdots & x_{2m}(b) \\ & \vdots & & & & & \vdots & & \\ x_{n_a1}(a) & x_{n_a2}(a) & \cdots & x_{n_am}(a) & & x_{n_b1}(b) & x_{n_b2}(b) & \cdots & x_{n_bm}(b) \end{matrix}$$

这是建立线性判别函数的基础数据。

（二）费歇尔准则下的判别函数

分别把总体 A、B 中取出的两组样品观测值代入线性判别函数式(5-1)，得两组判别函数值：

$$y_i(a) = \sum_{j=1}^{m} c_j x_{ij}(a) \qquad (i = 1,2,\cdots,n_a)$$

$$y_k(b) = \sum_{j=1}^{m} c_j x_{kj}(b) \qquad (k = 1,2,\cdots,n_b)$$

令

$$Q = [\bar{y}(a) - \bar{y}(b)]^2 \tag{5-2}$$

$$H = \sum_{i=1}^{n_a}[y_i(a) - \bar{y}(a)]^2 + \sum_{k=1}^{n_b}[y_k(b) - \bar{y}(b)]^2 \tag{5-3}$$

式中

$$\bar{y}(a) = \frac{1}{n_a}\sum_{i=1}^{n_a}\sum_{j=1}^{m} c_j x_{ij}(a) = \sum_{j=1}^{m} c_j \frac{1}{n_a}\sum_{i=1}^{n_a} x_{ij}(a) = \sum_{j=1}^{m} c_j \bar{x}_j(a)$$

$$\bar{y}(b) = \frac{1}{n_b}\sum_{k=1}^{n_b}\sum_{j=1}^{m} c_j x_{kj}(b) = \sum_{j=1}^{m} c_j \bar{x}_j(b)$$

确定判别函数的准则是要求 Q 达到最大，而 H 达到最小。这一准则是1963年由费歇尔(Fisher)提出的，故称作费歇尔准则。

若满足 H 最小，说明两组样品判别函数值的离散度最小，即样品点在平面

$$y = \sum_{j=1}^{m} c_j x_j$$

上的分布比在任何其他一个平面上的分布都要集中。Q 最大，表明两组样品点的中心距离最大。满足上述两个条件求得的线性判别函数能够最大限度地区分开 A、B 两个总体。

要求 Q 最大，H 最小，则等价于要求

$$V = Q/H$$

达到最大。V 为 $c_j(j=1,2,\cdots,m)$ 的函数，令 V 对 c_j 的偏导数等于0，得如下方程组：

$$\frac{\partial V}{\partial x_j} = 0 \qquad (j = 1,2\cdots,m) \tag{5-4}$$

对式(5-4)化简整理，则有

$$\sum_{k=1}^{m} s_{jk} c_k = d_j \qquad (j = 1,2,\cdots,m) \tag{5-5}$$

式中

$$s_{jk} = \sum_{i=1}^{n_a}[x_{ij}(a) - \bar{x}_j(a)][x_{ik}(a) - \bar{x}_k(a)]$$
$$+ \sum_{i=1}^{n_b}[x_{ij}(b) - \bar{x}_j(b)][x_{ik}(b) - x_k(b)] \quad (j、k = 1,2,\cdots,m)$$

$$d_j = [\bar{x}_j(a) - \bar{x}_j(b)] \qquad (j = 1,2,\cdots,m)$$

由线性方程组(5-5)可解出判别系数 c_j,得线性判别函数式(5-1)。

式(5-5)的矩阵形式为:

$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \cdot & \cdot & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \end{bmatrix}$$

因而判别系数

$$C = S^{-1}D = \left[\sum_{i=1}^{m} s_{1i}^{-1} d_i \sum_{i=1}^{m} s_{2i}^{-1} d_i \cdots \sum_{i=1}^{m} s_{mi}^{-1} d_i \right]' \qquad (5\text{-}6)$$

式中 s_{ij}^{-1} 是 S 的逆矩阵 S^{-1} 中第 i 行第 j 列的元素。

三、显著性检验及判别指数

(一)显著性检验

在两总体判别分析中,是假定两组样品取自两个不同的总体。如果两组样品的 m 个变量在统计上没有明显的差异,那么由样品观测值建立的判别函数就没有意义。为此,要对两个总体进行显著性检验。这里给出两种常用的检验方法。

1. *正判率法*

利用线性判别函数对已知总体的 $N(n_a+n_b)$ 个样品的总体重新判别,若判对了 $n(n \leqslant N)$ 个,定义 $R=n/N$ 为正判率。R 越大,两个总体的差异就越明显。

2. *F 分布检验法*

假设 $H_0: d_1=d_2=\cdots=d_m=0$

在假设 H_0 为真的条件下,统计量

$$F = \left(\frac{n_a n_b}{(n_a + n_b)(n_a + n_b - 2)} \right) \cdot \left(\frac{n_a + n_b - m - 1}{m} \right) D^2 \qquad (5\text{-}7)$$

服从第一自由度为 m,第二自由度为(n_a+n_b-m-1)的 $F(m, n_a+a_b-m-1)$ 分布。式(5-7)中 D^2 为马哈拉诺比斯(Mahalanobis)距离,即

$$D^2 = (n_a + n_b - 2) \sum_{i=1}^{m} \sum_{j=1}^{m} d_i d_j s_{ij}^{-1} = (n_a + n_b - 2) \sum_{j=1}^{m} c_j d_j$$

对于给定的检验水平 α,查 F 分布表得 F_α,若由式(5-7)计算的 $F > F_\alpha$,则认为假设 H_0 不成立,即两个总体差异显著。

(二)判别指数

在假设检验显著的条件下,定义

$$y_c = \frac{n_a \bar{y}(a) + n_b \bar{y}(b)}{n_a + n_b} \qquad (5\text{-}8)$$

为判别样品总体的判别指数。若 $\bar{y}(a) > y_c > \bar{y}(b)$,把待判样品 X 的观测值 $x_1, x_2, \cdots x_m$ 代入线性判别函数,得判别函数值

$$y = \sum_{j=1}^{m} c_j x_j$$

当 $y > y_c$ 时,样品 X 的总体为 A,否则为 B。

§2 多总体判别分析

一、原始数据

如果从 $G(G>2)$ 个总体 a_1、$a_2\cdots a_G$ 中分别取出 n_1、n_2、$\cdots$、n_G 个样品，并且每个样品有 m 个变量，那么样品构成的观测样本为

$$X_{gk}=\begin{bmatrix} x_{gk}^{(1)} \\ x_{gk}^{(2)} \\ \vdots \\ x_{gk}^{(m)} \end{bmatrix} \quad (g=1,2,\cdots,G;k=1,2,\cdots,n_g)$$

式中 $x_{gk}^{(i)}$ 为总体 $a_g(g=1,2,\cdots,G)$ 中第 $k(k=1,2,\cdots,n_k)$ 个样品第 $i(i=1,2,\cdots,m)$ 个变量的观测值。

二、多总体判别分析的判别函数

如果把取出的 G 组样品视为 G 个总体，并记为

$$A=(a_1,a_2,\cdots,a_G)$$

那么对于待判别的一个样品 $X(X\in A)$ 来说，在对它所属的总体作出判定之前，它属于任何一个总体都是可能的，只是归属总体 $a_g(g=1,2,\cdots,G)$ 的概率不同。如果把 $a_1,a_2,\cdots,a_G$ 视为总体样本空间的一个划分，那么由 Bayes 公式可以求得样品 X 属于 $a_g(g=1,2,\cdots,G)$ 的条件（后验）概率：

$$\begin{aligned} P(a_g/X) &= P(a_g)P(X/a_g)\Big[\sum_{j=1}^{G}P(a_j)P(X/a_j)\Big]^{-1} \\ &= P_g f_g(X)\Big[\sum_{j=1}^{G}P_j f_j(X)\Big]^{-1} \end{aligned} \tag{5-9}$$

式(5-9)中的 P_g，$f_g(X)$ 分别是总体 a_g 的先验概率和概率密度。

依据条件概率 $P(a_g/X)$ 的相对大小，则可对未知样品 X 的总体作出判断。若 $P(a_k/X)$ 是条件概率中的最大者，那么把未知样品 X 的总体判定为 a_k，判错的概率就最小。在计算条件概率时，式(5-9)的分母是一个与 g 无关的常量 C，若取式(5-9)的分子，记为

$$E_g(X)=P_g f_g(X) \quad (g=1,2,\cdots,G) \tag{5-10}$$

那么式(5-10)的函数值仅是条件概率 $P(a_g/X)$ 的 C 倍，因此按 $E_g(X)$ 函数值的相对大小判定未知样品 X 的总体与式(5-9)是等价的。式(5-10)是多总体判别的一般判别函数。

三、正态总体的判别函数

不同分布的总体具有不同的先验概率和概率密度，因此要想利用式(5-10)确定样品 X 的归属，还要具体确定式中的 P_g 和 $f_g(X)$。

设 $a_g(g=1,2,\cdots,G)$ 是以期望向量 μ_g 和相同协方差距阵 $\sum$ 为参数的正态总体，那么 a_g 的概率密度为：

$$f_g(X)=\frac{\left|\sum^{-1}\right|^{1/2}}{(2\pi)m/2}\exp\Big[-\frac{1}{2}(X-\mu_g)'\sum^{-1}(X-\mu_g)\Big] \tag{5-11}$$

根据样品构成的观测样本可求得式(5-11)中期望向量 μ_g 和协方差矩阵 $\sum$ 的估计值 $\overline{X}_g$ 和 S，并记矩阵 S 的递矩阵为 S^{-1}。

观测样本的期望向量

$$\overline{X}_g = \begin{bmatrix} \overline{x}_g^{(1)} \\ \overline{x}_g^{(2)} \\ \vdots \\ \overline{x}_g^{(m)} \end{bmatrix} \quad (g = 1,2,\cdots,G)$$

式中

$$\overline{x}_g^{(i)} = \frac{1}{n_g}\sum_{k=1}^{n_g} x_{gk}^{(i)} \quad (i = 1,2,\cdots,m)$$

观测样本的协方差距阵

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \cdot & \cdot & \cdots & \cdot \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{bmatrix}$$

式中

$$s_{ij} = \frac{1}{N-G}\sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_{gk}^{(i)} - \overline{x}_g^{(i)})(x_{gk}^{(j)} - \overline{x}_g^{(j)})$$

$$(i,j = 1,2,\cdots,m; N = n_1 + n_2 + \cdots + n_G)$$

由此，可把式(5-11)近似写为：

$$f_g(X) = \frac{|S^{-1}|^{1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(X - \overline{X}_g)'S^{-1}(X - \overline{X}_g)\right] \tag{5-12}$$

其中 $X=(x^{(1)}x^{(2)}\cdots x^{(m)})'$。

把式(5-12)和 $P_g=n_g/N$ 代入式(5-10)得正态总体假设下的判别函数

$$E_g(X) = P_g f_g(X) = \frac{n_g}{N}\cdot\frac{|S^{-1}|^{1/2}}{(2\pi)^{m/2}}\exp\left[-\frac{1}{2}(X - \overline{X}_g)'S^{-1}(X - \overline{X}_g)\right] \tag{5-13}$$

$$(g = 1,2,\cdots,G)$$

在对样品 X 进行判别时，需要的仅是使 $E_g(X)$ 为最大的 g，因此可对式(5-13)作进一步化简：

对式(5-13)两边取自然对数得

$$\ln(E_g(X)) = \ln\frac{n_g}{N} + \ln\frac{|S^{-1}|^{1/2}}{(2\pi)^{m/2}} - \frac{1}{2}X'S^{-1}X + X'S^{-1}\overline{X}_g - \frac{1}{2}\overline{X}_g'S^{-1}\overline{X}_g$$

去掉与 g 无关的项，并把 n_g/N 记为 q_g，得函数

$$F_g(X) = \ln q_g + X'S^{-1}\overline{X}_g - \frac{1}{2}\overline{X}_g'S^{-1}\overline{X}_g$$

记

$$C_g = S^{-1}\overline{X}_g = (c_g^{(1)}c_g^{(2)}\cdots c_g^{(m)})' \quad (g = 1,2,\cdots,G)$$

$$C_g^{(k)} = \sum_{l=1}^{m} s_{kl}^{-1}\overline{x}_g^{(l)} \quad (k = 1,2,\cdots,m)$$

则正态总体的判别函数为：

$$F_g(X) = \ln q_g + \sum_{k=1}^{m} c_g^{(k)}x^{(k)} - \frac{1}{2}\sum_{k=1}^{m} c_g^{(k)}\cdot\overline{x}_g^{(k)}$$

$$= \ln q_g + \sum_{k=1}^{m} c_g^{(k)}x^{(k)} + c_{0g} \tag{5-14}$$

$$(g = 1,2,\cdots,G)$$

四、对样品所属总体的判别

把样品 X 的观测值 $X=(x^{(1)}x^{(2)}\cdots x^{(m)})'$ 代人 $F_g(X)$，得 $F_1(X),F_2(X),\cdots,F_G(X)$，若

$$F_k(X)=\max_{1\leqslant g\leqslant G}\{F_g(X)\}$$

则判定样品 X 的总体为 a_k，它属于总体 a_k 的条件概率为

$$P_k=\exp[F_k(X)]/\sum_{j=1}^{G}\exp[F_j(X)] \tag{5-15}$$

$$(k=1,2,\cdots,G)$$

五、多总体判别分析的计算步骤

在总体服从正态分布的条件下，多总体判别分析的计算步骤是：

1. 计算各总体样本的变量平均值

$$\bar{x}_g^{(i)}=\frac{1}{n_g}\sum_{k=1}^{n_g}x_{gk}^{(i)}\quad(g=1,2,\cdots,G;i=1,2,\cdots,m)$$

2. 计算协方差距阵的逆矩阵

由原始数据求得 $\sum$ 的估计值 $S=[s_{ij}]_{m\times m}$

$$s_{ij}=\frac{1}{N-G}\sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_g^{(i)}-\bar{x}_{gk}^{(i)})(x_{gk}^{(j)}-\bar{x}_g^{(j)})\qquad(i,j=1,2,\cdots,m)$$

而 S 的递矩阵

$$S^{-1}=[s_{ij}^{-1}]_{m\times m}$$

3. 判别函数

$$F_g(X)=\ln q_g+\sum_{k=1}^{m}c_g^{(k)}x^{(k)}-\frac{1}{2}\sum_{k=1}^{m}c_g^{(k)}\bar{x}_g^{(k)}$$

$$=\ln q_g+\sum_{k=1}^{m}c_g^{(k)}x^{(k)}+c_{og}$$

$$(g=1,2,\cdots,G)$$

式中

$$c_g^{(k)}=\sum_{l=1}^{m}s_{kl}^{-1}\bar{x}_g^{(l)},\ c_{og}=-\frac{1}{2}\sum_{k=1}^{m}c_g^{(k)}\cdot\bar{x}_g^{(k)}$$

六、判别函数的检验

在实际的地质研究工作中，通常认为总体之间的差异是显著的。但是对已知的来自 G 个总体的 N 个样品进行判别验正后，出现样品被判定的总体与原来所属的总体有差异，甚至会出现较大的差异。出现这种情况的原因可能有两个方面：一是样品原来所属的总体是否正确；另一方面是所选取的变量不能充分表现各总体之产间的差异。这里仅就变量的区分能力进行检验。

1. 正判率法

利用判别函数对已知总体的 N 个样品的总体重新进行判别，若判对了 n 个，那么称 $R=n/N$ 为正判率。R 越接近 1，判别函数的判别效果越好。

2. 马哈拉诺比斯距离 D^2 检验法。

$$D^2=\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{G}n_k s_{ij}^{-1}(\bar{x}_k^{(i)}-\bar{x}^{(i)})(\bar{x}_k^{(j)}-\bar{x}^{(j)}) \tag{5-16}$$

式中

$$\bar{x}^{(i)}=\frac{1}{N}\sum_{k=1}^{G}\sum_{j=1}^{n_k}x_{kj}^{(i)}=\frac{1}{N}\sum_{K=1}^{G}n_k\cdot\bar{x}_k^{(i)}\quad(i=1,2,\cdots,m)$$

统计量 D^2 服从自由度为 $m(G-1)$ 的 χ^2 分布，查表可定出 D^2 的临界值 D^*。当式(5-16)计算出的 D^2 大于 D^* 时，可以认为这 m 个变量能够区分开 G 个总体；否则，就认为所拟定的这 m 个变量不能对样品的归属作出正确的判别。在后一种情况下，应剔除一些不重的变量或者引进一些新的更有效的变量，重新建立判别函数。

§3 逐步判别分析

地质研究人员总希望能用尽可能少的变量就能解决需要判别的问题，也就是说，应当选择少数的有效变量进行判别分析。多余的变量参加判别时，不仅会使计算工作量增加，而且还有可能因相关变量的增加而导致求解判别函数的困难，因而，自然会产生类似逐步回归分析的想法，即对变量按其判别能力的大小，在计算过程中有进有出，从而保留那些对判别总体起决定作用的变量，剔除那些对判别总体作用小甚至不起作用的变量。

一、逐步判别方法原理

1. Wilks Λ 统计量

我们仍然用上一节的符号，把样本数据记为 X_{gk}，它来自 G 个具有相同协方差距阵的正态总体 $N(\mu_i, \sum)$。

现在定义样本内离差距阵 W、样本间离差矩阵 B 和总离差距阵 T 如下：

$$\bar{x}_g^{(i)} = \frac{1}{n_g}\sum_{k=1}^{n_g} x_{gk}^{(i)}$$

$$\bar{x}^{(i)} = \sum_{g=1}^{G}\sum_{k=1}^{n_g} x_{gk}^{(i)} \Big/ \sum_{g=1}^{G} n_g = \sum_{g=1}^{G} n_g \bar{x}_g^{(i)} \Big/ \sum_{g=1}^{G} n_g$$

$$w_{ij} = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_{gk}^{(i)} - \bar{x}_g^{(i)})(x_{gk}^{(j)} - \bar{x}_g^{(j)})$$

$$b_{ij} = \sum_{g=1}^{G} n_g(\bar{x}_g^{(i)} - \bar{x}^{(i)})(\bar{x}_g^{(j)} - \bar{x}^{(j)})$$

$$t_{ij} = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(x_{gk}^{(i)} - \bar{x}^{(i)})(x_{gk}^{(j)} - \bar{x}^{(j)})$$

$$(i,j = 1,2,\cdots,m)$$

$$W = [w_{ij}]_{m\times m},\ B = [b_{ij}]_{m\times m},\ T = [t_{ij}]_{m\times m}$$

可以证明

$$T = W + B$$

Wilks Λ 统计量

$$U = \frac{|W|}{|T|}$$

是在假设

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_G$$

下检验 m 个变量综合判别能力的统计量。它是两个行列式之比，U 越小，总体内部差异越小，各总体之间的差异就越明显。

2. “引入”与“剔除”变量的统计量

U 是检验变量综合判别能力的一个指标。现在就从 U 出发，找出检验某个变量 $x^{(r)}$ 判别能

力的 Wilks Λ 统计量。

对于任意一个 n 阶行列式 $|D|$，如果按列号 $r_1, r_2, \cdots, r_n$ 的顺序进行消去计算，那么行列式 $|D|$ 可以写成：

$$|D| = d_{r_1r_1}^{(0)} d_{r_2r_2}^{(1)} \cdots d_{r_nr_n}^{n-1} \quad (r_p \text{ 不相等}; p = 1, 2, \cdots, n)$$

由此可得，Wilks Λ 统计量

$$U = \frac{w_{r_1r_1}^{(0)} w_{r_2r_2}^{(1)} \cdots w_{r_mr_m}^{(m-1)}}{t_{r_1r_1}^{(0)} t_{r_2r_2}^{(1)} \cdots t_{r_mr_m}^{(m-1)}}$$

为了表明消列顺序，把 Wilks Λ 统计量改写为

$$U_{r_1r_2\cdots r_m} = \frac{w_{r_1r_1}^{(0)} w_{r_2r_2}^{(1)} \cdots w_{r_mr_m}^{(m-1)}}{t_{r_1r_1}^{(0)} t_{r_2r_2}^{(1)} \cdots t_{r_mr_m}^{(m-1)}} \tag{5-17}$$

由式(5-17)可导出引入或剔除变量 $x^{(r)}$ 的 Wilks Λ 统计量。

(1) "引入"变量 $x^{(r)}$ 的 Wilks Λ 统计量

设逐步判别分析进行了 P 步，共引入了 P 个(即前 P 步没有剔除变量)变量，记为

$$x^{(r_1)}, x^{(r_2)}, \cdots, x^{(r_p)}$$

根据式(5-17)有

$$U_{r_1r_2\cdots r_p} = \frac{w_{r_1r_1}^{(0)} w_{r_2r_2}^{(1)} \cdots w_{r_pr_p}^{(p-1)}}{t_{r_1r_1}^{(0)} t_{r_2r_2}^{(1)} \cdots t_{r_pr_p}^{(p-1)}} \tag{5-18}$$

在引入变量 $x^{(r_1)}, x^{(r_2)}, \cdots, x^{(r_p)}$ 之后，若再引入变量 $x^{(r)}$，则有

$$U_{r_1r_2\cdots r_pr} = \frac{w_{r_1r_1}^{(0)} w_{r_2r_2}^{(1)} \cdots w_{r_pr_p}^{(p-1)} w_{rr}^{(p)}}{t_{r_1r_1}^{(0)} t_{r_2r_2}^{(1)} \cdots t_{r_pr_p}^{(p-1)} t_{rr}^{(p)}} \tag{5-19}$$

由式(5-18)和式(5-19)可知，$w_{rr}^{(p)}/t_{rr}^{(p)}$ 是引入变量 $x^{(r)}$ 后，U 的改变因子，记为

$$U_r = w_{rr}^{(p)}/t_{rr}^{(p)} \quad (r \neq r_1, r_2, \cdots, r_p) \tag{5-20}$$

U_r 越小，说明变量 $x^{(r)}$ 在总体之间的差异越明显，它所起的判别作用就越大。因此，U_r 是检验变量 $x^{(r)}$ 判别能力的 Wilks Λ 统计量。记

$$F_1 = \frac{(1 - U_r)/(G - 1)}{U_r/(N - G - P)} = \frac{(N - G - P)(t_{rr}^{(p)} - w_{rr}^{(p)})}{(G - 1)w_{rr}^{(p)}} \tag{5-21}$$

这里 $N = n_1 + n_2 + \cdots + n_G$。

F_1 服从第一自由度为 $(G-1)$，第二自由度为 $(N-G-P)$ 的 F 分布。对于给定的检验水平 α，查 $F_\alpha(G-1, N-G-P)$ 分布表，得临界值 F_α。若由式(5-21)计算的 $F_1 > F_\alpha$，变量 $x^{(r)}$ 的判别能力强，把它引入判别函数。

(2) "剔除"变量 $x^{(r)}$ 的 Wilks Λ 统计量

设逐步判别分析进行了 P 步，共引入了 P 个变量 $x^{(r_1)}, x^{(r_2)}, \cdots, x^{(r_p)}$。它的第 $p+1$ 步拟剔除变量 $x^{(r)}(r \in (r_1, r_2, \cdots, r_p))$，此时，变量 $x^{(r)}$ 的判别能力可视为第 p 步引入 $x^{(r)}$ 的判别能力，即

$$\dot{U}_r = w_{rr}^{(p-1)}/t_{rr}^{(p-1)} \quad (r \in (r_1, r_2, \cdots, r_p)) \tag{5-22}$$

统计量

$$F_2 = \frac{(1 - \dot{U}_r/(G - 1)}{\dot{U}_r/(N - G - p + 1)} = \frac{(N - G - p + 1)(1 - \dot{U}_r)}{(G - 1)\dot{U}_r} \tag{5-23}$$

服从自由度为 $(G-1)$ 和 $(N-G-p+1)$ 的 F 分布。对于给定的检验水平 α，查 $F_\alpha(G-1, N-G$

$-p+1$)分布表，得临界值 F'_a。若 $F_2 \leqslant F'_a$，从判别函数中剔除变量 $x^{(r)}$，否则，把它保留在判别函数中。

3. 逐步判别分析的变换公式

逐步判别分析建立判别函数的过程与逐步回归建立回归方程的过程相似，不同之处仅是逐步判别分析要对 W 和 T 两个矩阵进行变换。

逐步判别分析的第 $p+1$ 步，不论是引入还是剔除变量 $x^{(r)}$，都是对 W 和 T 矩阵进行一次变换。第 $p+1$ 步消去第 r 列的变换公式如下：

$$w_{kl}^{(p+1)} = \begin{cases} 1/w_{rr}^{(p)} & (k=r, l=r) \\ w_{rl}^{(p)}/w_{rr}^{(p)} & (k=r, l\neq r) \\ -w_{kr}^{(p)}/w_{rr}^{(p)} & (k\neq r, l=r) \\ w_{kl}^{(p)} - \omega_{kr}^{(p)} \cdot w_{rl}^{(p)}/w_{rr}^{(p)} & (k\neq r, l\neq r) \end{cases} \tag{5-24}$$

$$t_{kl}^{(p+1)} = \begin{cases} 1/t_{rr}^{(p)} & (k=r, l=r) \\ t_{rl}^{(p)}/t_{rr}^{(p)} & (k=r, l\neq r) \\ -t_{kr}^{(p)}/t_{rr}^{(p)} & (k\neq r, l=r) \\ t_{kl}^{(p)} - t_{kr}^{(p)} \cdot t_{rl}^{(p)}/t_{rr}^{(p)} & (k\neq r, l\neq r) \end{cases} \tag{5-25}$$

二、判别函数的系数和对样品的判别

1. 判别函数的系数

设逐步判别进行了 p 步，共引入了 $v(v \leqslant m)$ 个变量，此时，按下式计算判别函数的系数。

$$\begin{cases} c_g^{(i)} = (N-G)\sum_{j\in v} w_{ij}^{(p)} \bar{x}_g^{(j)}, i \in v & (g=1,2,\cdots,G) \\ c_{og} = -\dfrac{1}{2}\sum_{i\in v} c_g^{(i)} \bar{x}_g^{(i)} & (g=1,2,\cdots,G) \end{cases} \tag{5-26}$$

2. 对样品的判别

假设待判别样品 $X=(x^{(1)}x^{(2)}\cdots x^{(m)})'$，那么它属于第 g 个总体的判别函数值为

$$F_g(X) = \ln q_g + \sum_{i\in v} c_g^{(i)} x^{(i)} + c_{og} \quad (g=1,2,\cdots,G)$$

若

$$F_r(X) = \max_{1\leqslant g\leqslant G}\{F_g(X)\}$$

那么样品归属于总体 a_r。它属于总体 a_r 的条件概率为

$$p_r = \exp[F_r(X)] / \sum_{j=1}^{G} \exp[F_j(X)]$$

§4 逐步判别分析 FORTRAN 源程序

本程序用于在正态总体假设下，根据给定的 F 统计检验量，从已知类别归属的多变量样本中，逐步选取适当数量的对判别归类起显著作用的变量建立判别未知样品归属的判别函数，并对已知归属的多变量样品进行验算，给出每个样品归属某类的后验概率及正判率。现将程序中的主要参数，符号及程序的使用方法说明如下：

一、主要参数及符号

1. 参数

n——整型变量，已知类别归属并参加建立判别函数的样品数；

m——整型变量，每个样品的变量数；

g——整型变量，样本数；

f1——引入变量的 F 统计检验量；

f2——剔除变量的 F 统计检验量；

l——选入判别函数的变量数。

2. 符号

x——存放原始数据二维数组名；

q——存放先验概率对数值的一维数组名；

w——存放类内离差的二维数组名；

t——存放总离差的二维数组名；

mx——存放类均值和总均值的二维实型数组名；

c——存放判别系数 $C_g^{(k)}$ 的二维数组名；

co——存放判别函数中 c_{og}项的一维数组名；

foi——第 i 个判别函数的常数项；

fik——第 i 个判别函数中第 k 个变量的系数。

3. 子程序

sdisc——逐步引入和剔除变量子程序；

sd1——判别归类子程序；

sd2——计算判别系数子程序；

sd3——计算任一样品的判别归类值及后验概率子程序。

二、程序使用说明

1. 原始数据文件

原始数据文件由已知类别归属的 n 个样品 m 个变量的观测值及样品所属总体号组成，其形式为：

$$x_{ij}^{(1)}, x_{ij}^{(2)}, \cdots, x_{ij}^{(m)}, r_i$$
$$(j = 1, 2, \cdots, n_i, i = 1, 2, \cdots, G)$$

其中 $x_{ij}^{(k)}$表示第 i 个总体中第 j 个样品第 k 个变量的观测值，而 r_i 为 $x_{ij}^{(k)}$所属的总体号。

2. 操作说明

在 DOS 操作系统下，键入逐步判别分析目的程序名 zbpb，具体操作步骤如下：

(1) 输入数据文件名(Input your data file name)；

(2) 输入样品总数 n、变量数 m 和总体数 g(Input n,m,g)；

(3) 输入引入和剔除变量的 F 统计检验量 f1 和 f2(Input f1 and f2)；

(4) 确定先验概率类型(To select a priori probabitity tepe)，并键入 pp 值：当各总体的样品数相等时，pp=1，即先验概率为 1/g；否则 pp=2，程序将自动计算先验概率，并建立判别函数和对已知样品进行判别验证。若判别函数中选入的变量数为 0，系统显示：是否改变 f1 或 f2 的值(To change f1 or f2 (y/n)?)键入 y，程序返回到输入 f1 和 f2 处，否则键入 n 或者判别函数中的变量数不等于 0 时、系统都将询问：是否想换一种先验概型(To change a priori Probabilitg tgpr (y/n)?)，回答 y，程序运行返回到先验概率选择处，否则键入 n，程序运行结束。

3. 主要输出结果

结果文件为 zbpb.dat，主要存储了三方面的内容：

(1) 基础数据

主要有引入和剔除变量的 F 检验统计量 f1 和 f2；先验概型 pp；类中样品数，类均值、总均值、类内离差矩阵和总离差矩阵。

(2) 中间结果

每一步计算涉及的变量号；判别函数中已引入的变量数；计算的 F 值和方差贡献 U_r 值；判别函数的常数项 foi 和系数 fik 等。若无变量引入，将输出信息 fail，表明 F 统计量给的不合适。

(3) 验算结果

对参与建立判别函数的 n 个样品重新判别归类，给出归类结果和后验概率。

三、源程序

1. 逐步判别分析流程

逐步判别分析流程如图 5-2 所示。

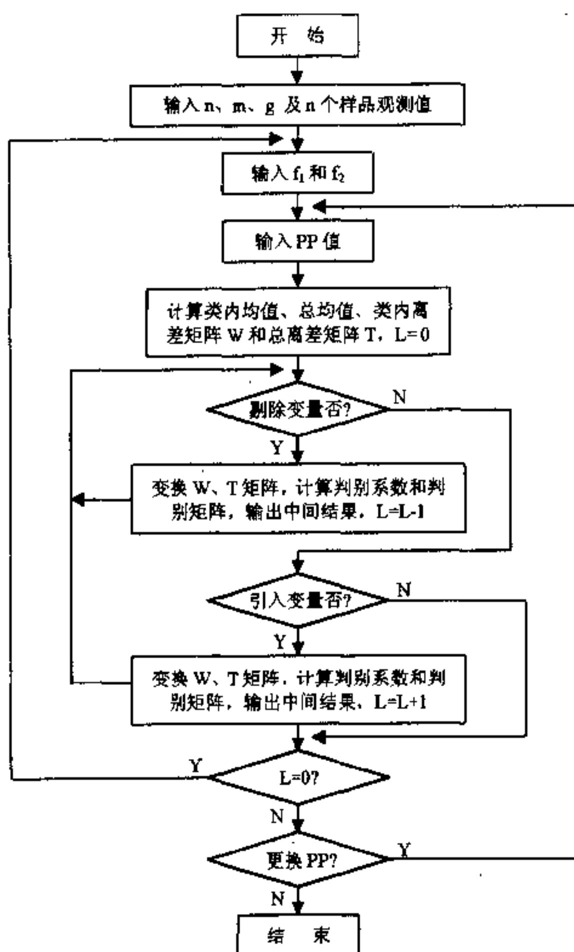


图 5-2 逐步判别分析流程图

2. **逐步判别分析 *FORTRAN* 源程序**

```
$debug
          program zbpb
c         This is a successive discriminant program
          common /cc/m,n,g,l,z(20),c(20,20),ng(50),co(20)
          common /s1/mx(51,20),w(20,20),d2ef(171),fef(171)
          common /s2/dm(51,51),qln(50),x(300,21),pr(50),fam
          integer g,dm,pp
          real mx
          character fname * 10,type,fail
          character * 10 fam
          write ( * ,'(a)') ' Successive discriminant procedure '
          fam='zbpb.dat '
          open(2,file=fam)
          write ( * ,'(a)') ' Input your data file name'
          read ( * ,'(a)') fname
```

```
      open (3,file=fname)
      write( * , * ) 'Input n,m,g '
      read ( * , * ) n,m,g
      do 10 i=1,1000
10    read (3, * ,end=20,err=15) (x(i,j),j=1,m+1)
15    write( * , * ) 'I= ',i
      write( * , * ) 'Err of file '
      stop
20    write( * , * ) 'End of file ',i-1
      nc=0
25    write ( * , * ) ' Input f1 and f2'
      read ( * , * ) f1,f2
      write(2,30) f1,f2
30    format(1x,'f1=',f10.4,5x,'f2=',f10.4)
40    write ( * , * ) ' To select a priori probability type'
      write ( * , * ) ' pp: 1,2,pp ='
      read ( * , * ) pp
      write(2,50) pp
50    format(1x,'pp=',i2)
      call sdisc(pp,f1,f2,fail)
      if(fail.eq.'y'.or.fail.eq.'Y') then
      write( * , * ) 'To change f1 or f2 (y/n)? '
      read( * ,'(a)') type
      if(type.eq.'y'.or.type.eq.'Y') go to 25
      end if
      write ( * , * ) 'To change a priori probability type (y/n)? '
      read ( * ,'(a)') type
      if(type.eq.'y'.or.type.eq.'Y') go to 40
      write ( * ,'(///)')
      stop
      end

      subroutine sdisc(pp,f1,f2,fail)
      dimension t(20,20)
      common /cc/m,n,g,l,z(20),c(20,20),ng(50),co(20)
      common /s1/mx(51,20),w(20,20),d2ef(171),fef(171)
      common /s2/dm(51,51),qln(50),x(300,21),pr(50),fam
      integer g,g1,step,h,r,r12,dm,pp
      character * 10 fam,fail
      real mx
```

```
      g1=g+1
      m1=m+1
      do 100 kh=1,g1
      do 100 i=1,m
100   mx(kh,i)=0.
      do 105 kh=1,g
105   ng(kh)=0
      do 115 k=1,n
      h=x(k,m1)
      ng(h)=ng(h)+1
      do 110 i=1,m
110   mx(h,i)=x(k,i)+mx(h,i)
115   continue
      do 125 i=1,m
      d=0.
      do 120 kh=1,g
      d=mx(kh,i)+d
120   mx(kh,i)=mx(kh,i)/ng(kh)
125   mx(g1,i)=d/n
      write (2,130)
130   format(' The sample number of cach class ')
      write (2,135) (ng(i),i=1,g)
135   format (10(2x,i3))
      write (2, * ) ' Mean values matrix'
      do 140 i=1,g1
140   write (2,145) (mx(i,j),j=1,m)
145   format (9(2x,f12.4))
      do 155 k=1,m
      do 150 i=1,m
      t(k,i)=0.
      w(k,i)=0.
150   continue
155   continue
      do 170 k=1,n
      do 165 i=1,m
      d=x(k,i)-mx(g1,i)
      z(i)=d
      do 160 j=1,i
      t(i,j)=d * z(j)+t(i,j)
160   continue
```

```
165       continue
170       continue
          do 185 kh=1,g
          d=ng(kh)
          do 180 i=1,m
          d1=mx(kh,i)-mx(g1,i)
          z(i)=d*d1
          do 175 j=1,i
          w(i,j)=d1*z(j)-w(i,j)
175       continue
180       continue
185       continue
          do 195 i=1,m
          do 190 j=1,i
          w(i,j)=t(i,j)-w(i,j)
          w(j,i)=w(i,j)
          t(j,i)=t(i,j)
190       continue
195       continue
          write (2,*) ' Matrix W '
          do 200 i=1,m
200       write (2,145) (w(i,j),j=1,m)
          write (2,*) ' Matrix T '
          do 205 i=1,m
205       write (2,145) (t(i,j),j=1,m)
          do 210 kh=1,g
          if(pp.eq.2) then
          d=(ng(kh)+0.)/n
          else if(pp.eq.1) then
          d=1./g
          end if
210       qln(kh)=alog(d)
          l=0
          eps=1.e-6
          u=1.
          do 215 i=1,m
215       z(i)=0.
          write (2,*) ' Iteration procedure '
          do 245 step =1,100
          umax=eps
```

```
      umin=10000.
      do 220 i=1,m
      if(z(i).eq.0.) then
      if(t(i,i).ge.eps) then
      ui=w(i,i)/t(i,i)
      if(ui.ge.eps) then
      if(ui.lt.umin) then
      umin=ui
      imin=i
      end if
      end if
      end if
      else
      ui=t(i,i)/w(i,i)
      if(ui.gt.umax) then
      umax=ui
      imax=i
      end if
      end if
220   continue
      f12=(1-umax)*(n-l-g+1)/(umax*(g-1))
      if (f12.le.f2) then
      l=l-1
      r=imax
      r12=-r
      z(r)=0.
      else
      f12=(1-umin)*(n-l-g)/(umin*(g-1))
      if (f12.le.f1+eps) then
      go to 250
      else
      l=l+1
      r=imin
      r12=r
      z(r)=1.
      end if
      end if
      u=w(r,r)/t(r,r)*u
      x2=-(n-1-(l+g)/2.)*alog(u)
      write (2,225) step,r12,l,f12,u,x2
```

```
225     format(/' Step=',i3/' x=',i2,' l=',i3,
     #  'f=',f10.6,' u=',f10.6,' ch12=',f10.6)
        d=1./t(r,r)
        t(r,r)=1.
        d1=1./w(r,r)
        w(r,r)=1.
        do 230 j=1,m
        t(r,j)=t(r,j)*d
        w(r,j)=w(r,j)*d1
230     continue
        do 240 i=1,m
        if (i.ne.r) then
        d=t(i,r)
        t(i,r)=0.
        d1=w(i,r)
        w(i,r)=0.
        do 235 j=1,m
        t(i,j)=t(i,j)-d*t(r,j)
        w(i,j)=w(i,j)-d1*w(r,j)
235     continue
        end if
240     continue
        call sd2
        call sd1(0)
245     continue
250     continue
        if(l.eq.0) then
        fail='y'
        write (2,*) 'fail'
        go to 260
        else
        fail='n'
        end if
        write (2,'(/a)') ' The results '
        write (2,255) l,u,x2
255     format(' l=',i3,' u= ',f10.6,' ch12=',f10.6)
        call sd2
        call sd1(1)
260     continue
        return
```

```
      end

      subroutine sd1(p)
      integer g,g1,p,dm
      common /cc/m,n,g,l,z(20),c(20,20),ng(50),co(20)
      common /s2/dm(51,51),qln(50),x(300,21),pr(50),fam
      character * 10 fam
      dimension xm(51)
      g1=g+1
      m1=m+1
      do 20 k=1,g1
      do 10 kh=1,g1
      dm(k,kh)=0
10    continue
20    continue
      do 40 k=1,n
      call sd3(k,kmax,ymax)
      kh=x(k,m1)
      dm(kmax,kh)=dm(kmax,kh)+1
      if(p.ne.0) write (2,30) k,kh,kmax,kmax,ymax
30    format (2x,i3,2x,'OClass:',i2,2x,
     #  'DClass:',i2,' P(',i1,'/x)=',f8.4)
      if(p.ne.0) write(2,30) kh,kmax,ymax
30    format(2x,i2,1x,i2,f6.4)
40    continue
      do 60 kh=1,g
      n1=0
      n2=0
      do 50 k=1,g
      n1=dm(kh,k)+n1
      n2=dm(k,kh)+n2
50    continue
      dm(kh,g1)=n1
      dm(g1,kh)=n2
60    continue
      dm(g1,g1)=n
      write (2,*) ' Discriminant matrix'
      do 70 i=1,g1
70    write (2,'(10(1x,i3))') (dm(i,j),j=1,g1)
      sm=0.
```

```
      do 80 i=1,g1-1
      xm(i)=dm(i,i)/float(dm(g1,i))
80    sm=sm+dm(i,i)
      xm(g1)=sm/dm(g1,g1)
      write(2,90) (xm(i),i=1,g1)
90    format(1x,f8.2,f6.2,8f6.2)
      return
      end

      subroutine sd2
      common /cc/m,n,g,l,z(20),c(20,20),ng(50),co(20)
      common /s1/mx(51,20),w(20,20),d2ef(171),fef(171)
      common /s2/dm(51,51),qln(50),x(300,21),pr(50),fam
      character * 10 fam
      integer g
      real mx
      nn=(g-1) * g/2
      do 30 kh=1,g
      coo=0.
      do 20 i=1,m
      ci=0.
      if(z(i).ne.0.) then
      do 10 j=1,m
      if(z(j).ne.0.) ci=w(i,j) * mx(kh,j)+ci
10    continue
      end if
      c(kh,i)=(n-g) * ci
      coo=ci * mx(kh,i)+coo
20    continue
      co(kh)=-coo * (n-g)/2
30    continue
      write (2,'(A)') ' Discriminant coefficient matrix c '
      do 40 i=1,g
40    write (2,45) co(i),(c(i,j),j=1,M)
45    format (7(2x,f15.4))
      do 65 i=1,g
      f0=co(i)+qln(i)
      write(2,50) i,f0
      do 60 j=1,m
      if(c(i,j).ne.0.) write(2,55) i,j,c(i,j)
```

```
50      format(1x,' f0',i1,'=',f15.4)
55      format(1x,' f',i1,i1,'=',f15.4)
60      continue
65      continue
        kd=0
        do 85 kh=2,g
        do 80 k=1,kh-1
        kd=kd+1
        d2=0.
        do 70 i=1,m
        if(z(i).ne.0.) then
        d2=(c(kh,i)-c(k,i))*(mx(kh,i))+d2
        end if
70      continue
        d2ef(kd)=d2
        fh=d2*(n-g-l+1)*ng(kh)*ng(k)
        fef(kd)=fh/(l*(n-g)*(ng(kh)+ng(k)))
80      continue
85      continue
        write (2,*) ' Matrix D2ef '
        j2=0
        do 90 i=2,g
        j1=j2-1
        j2=(i-1)*i/2
90      write (2,15) (d2ef(j),j=j1,j2)
        write (2,*) ' Matrix Fef '
        j2=0
        do 95 i=2,g
        j1=j2+1
        j2=(i-1)*i/2
95      write (2,15) (fef(j),j=j1,j2)
        return
        end

        subroutine sd3(k,kmax,ymax)
        common /cc/m,n,g,l,z(20),c(20,20),ng(50),co(20)
        common /s2/dm(51,51),qln(50),x(300,21),pr(50),fam
        character*10 fam
        integer g
        ymax=-10000.
```

```
      do 40 kh=1,g
      yh=qln(kh)+co(kh)
      do 20 i=1,m
      if(z(i).ne.0.) yh=c(kh,i)*x(k,i)+yh
20    continue
      pr(kh)=yh
      if(yh.gt.ymax) then
      ymax=yh
      kmax=kh
      end if
40    continue
      d=0.
      do 60 kh=1,g
      e=pr(kh)-ymax
      pr(kh)=exp(e)
60    d=d+pr(kh)
      ymax=pr(kmax)/d
      return
      end
```

§5 应用算例

判别分析在地质研究工作中有着广泛的应用，下面列举几个应用算例。

【例 1】 确定地层界线

我国辽宁前震旦系，由于混合岩化作用，使炮台山砾岩上、下的千枚岩用肉眼及镜下都难以分辨，由此导致地层对比的分歧意见：其一认为前震旦系为连续沉积，中间没有大的沉积间断；其二认为前震旦系分为两个沉积旋回，界线为砾岩上千枚岩。中间炮台山砾岩和其下千枚岩为太古代鞍山群的樱桃园组，而砾岩之上千枚岩为太古代辽河群的浪子山组。

为统一认识，在地质上可靠的樱桃园组顶部及浪子山组底部分别取了 15、25 块岩样，分析了它们的元素 TFe、Fe_2O_3、FeO、SiO_2、Al_2O_3、MgO、CaO、k_2O、Na_2O、MnO、TiO。利用逐步判别分析，引入 FeO、CaO、MgO 和 MnO 四个变量，得浪子山组底、樱桃园组顶千枚岩判别函数 $F_1(X)$ 和 $F_2(X)$ 如下：

$$F_1(X)=6.6029\mathrm{FeO}-0.7159\,\mathrm{CaO}+5.9735\,\mathrm{MgO}-11.5155\,\mathrm{MnO}-29.1880$$

$$F_2(X)=25.3260\mathrm{FeO}-21.5900\,\mathrm{CaO}-11.9730\,\mathrm{MgO}-1.5350\,\mathrm{MnO}-27.278$$。

利用以上两个判别函数，对砾岩样品进行判别，结果都判为樱桃园组的千枚岩。这说明砾岩上、下的千枚岩差异显著，砾岩的成分与其下千枚岩相似，故砾岩与其上千枚岩之间为前震旦系两大沉积旋回的分界线。

【例 2】 生油岩热演化阶段的判别

为了确定生油岩的热演化阶段，表 5-1 中列出了我国有关探区的 66 个生油岩的地层年龄 (t)、现今温度 (T) 以及埋藏深度 (H)。这些生油岩处于未成熟、成熟、高成熟和过成熟四个演化

阶段。

共设计六个变量：$T+273$、t、H、$1/H$、$\ln(T+273)$、$1/(T+273)$，取 $f_1=f_2=0$，引入四个变量，得生油岩演化阶段判别函数式如下：

$$F_1 = -431.6772x_1 + 4.3990x_2 - 0.2610x_3 + 200298.2x_5 - 510438.9$$

$$F_2 = -432.6836x_1 + 4.4043x_2 - 0.2610x_3 + 200782.5x_5 - 512924.1$$

$$F_3 = -433.8340x_1 + 4.4108x_2 - 0.2608x_3 + 201345.2x_5 - 515827.7$$

$$F_4 = -434.4465x_1 + 4.4140x_2 - 0.2598x_3 + 201681.6x_5 - 517606.3$$

判别函数中变量的引入顺序及各演化阶段的正判率分别见表 5-2 和表 5-3。

表 5-1　我国生油岩的演化参数表

演化阶段	序　号	地　区	热演化参数		
			T/℃+273	t/Ma	H/m
未成熟	1	松辽盆地（青$_{2+3}$）	337	125.005	1 000
	2	松辽盆地（青$_1$）	328	123	1 000
	3	松辽盆地（姚$_2$）	334	125	1 750
	4	岐口凹陷	347	33.75	1 680
	5	泌阳凹陷	342	27.999	1 460
	6	湖北（二叠系）	338	242	3 200
	7	潜江凹陷	343	34.33	1 600
	8	高邮凹陷	348	15	1 580
	9	惠民凹陷	351	8.04	1 350
	10	沾化凹陷	344	10.76	1 600
	11	东明凹陷	352	8.5	2 250
	12	松辽盆地（姚$_2$）	228.5	127	1 480
成熟	13	松辽盆地（青$_{2+3}$）	341	127	1 550
	14	松辽盆地（青$_{2+1}$）	367	127	2 199
	15	松辽盆地（青）	362	124.2	1 799
	16	松辽盆地（姚$_2$）	364	120.4	1 851
	17	松辽盆地（姚$_2$）	370	120.4	1 970
	18	岐口盆地	359	35	2 001
	19	泌阳凹陷	353	31	1 700
	20	泌阳凹陷	367	31	2 099
	21	辽河陷陷	364	48.5	2 001
	22	东台凹陷（阜宁组）	364	34	2 201
	23	湖北（二叠系）	348	257.6	3 450
	24	高邮凹陷（集宁组）	356	17.5	2 200
	25	高邮凹陷（集宁组）	360	18	2 700
	26	沾化凹陷	367	10.7	2 300
	27	沾化凹陷	364	12.8	2 200
	28	沾化凹陷	368	13	2 500
	29	东明凹陷	362	11.6	3 500

演化阶段	序号	地区	热演化参数		
			T/C+273	t/Ma	H/m
高成熟	30	松辽盆地(青$_{2+3}$)	380	129	3 100
	31	松辽盆地(青$_1$)	377	127.88	2 350
	32	松辽盆地(青$_1$)	402	127.88	2 499
	33	松辽盆地(姚$_2$)	383	124.64	2 299
	34	松辽盆地(姚$_2$)	377	121.80	2 150
	35	松辽盆地(姚$_2$)	389	125.008	2 401
	36	岐口凹陷	419	36	3 701
	37	湖北(二叠系)	392	271.25	4 050
	38	泌阳凹陷	376	34.01	2 300
	39	泌阳凹陷	392	34.05	2 799
	40	东台凹陷(阜宁组)	410	54.25	3 201
	41	东台凹陷(阜宁组)	398	37.005	2 900
	42	东台凹陷(阜宁组)	403	35.5	3 701
	43	湖北(二叠系)	399	271.25	4 001
	44	高邮凹陷(集宁组)	380	23.5	3 000
	45	高邮凹陷(集宁组)	382	24	3 400
	46	惠民凹陷	383	13	2 800
	47	惠民凹陷	386	13.5	3 100
	48	高邮凹陷(集宁组)	399	25	4 100
	49	沾化凹陷	381	16	2 700
	50	沾化凹陷	385	16.5	3 000
	51	沾化凹陷	387	16.5	3 100
	52	东明凹陷	410	15.05	3 550
	53	东明凹陷	412	18	3 550
	54	松辽盆地(青$_{2+3}$	400	130	3 400
过成熟	55	松辽盆地(姚$_2$)	434	123.205	3 555
	56	江汉盆地(潜江组)	440	37.005	4 300
	57	东台凹陷(阜宁组)	444	60.005	4 100
	58	东台凹陷(阜宁组)	433	57.005	5 100
	59	湖北(二叠系)	469	285	5 701
	60	湖北(二叠系)	572	285.5	7 199
	61	高邮凹陷(集宁组)	403	26	4 001
	62	东明凹陷	435	20	3 700
	53	东明凹陷	422	18	3 800
	64	泌阳凹陷	433	36.05	4 140
	65	湖北(二叠系)	454	271.29	5 200
	66	湖北(二叠系)	440	271.29	4 900

表 5-2　变量引入顺序及 F 检验量

引入顺序	变量号	变　量	F 检验量
1	x_4	$\ln(T+273)$	90.107 8
2	x_1	$T+273$	29.072 5
3	x_2	t	0.433 9
4	x_5	H	0.297 9

表 5-3　生油岩演化阶段正判率

演化阶段	正判率/%	演化阶段	正判率/%
未成熟	83	高成熟	96
成　熟	91	过成熟	92

各演化阶段的正判率都超过了 80%，故可把上述判别函数用于判别生油岩演化阶段。如珠江口盆地第三系生油岩为中新世至晚渐新世，地层绝对年龄约为 16～30 百万年，埋藏深度为 2 200 米，地层温度 104 ℃。按上述判别函数式计算，绝对年龄以 25 百万年计，判别函数值为

$$F_1 = 514572.6$$

$$F_2 = 514581.3$$

$$F_3 = 514582.5$$

$$F_4 = 514570.8$$

最大值为 F_3，判定属于高成熟阶段。

【例 3】 岩性剖面的反演

测井参数是岩石的效应，其观测值的差异主要取决于岩性，即决定于组成岩石的矿物成分、颗粒的大小、结构和岩石孔隙中所含流体的性质。也就是说，根据钻孔的测井参数，可以反演出钻孔的岩性剖面。

某砂砾岩油田，岩心实物少，但测井资料较丰富。为开展砾岩油藏描述研究工作，利用多数井共有的测井参数：

微电极 2(x_1)、2.5 m 梯度(x_2)、4 m 梯度(x_3)、感应电导(x_4)、声波(x_5)、浅测向(x_6)、补偿中子(x_7)、井经(x_8)、微电极差(x_9)反演了 30 余口井的岩性剖面，为沉积相的纵向分布和横向展布及储层研究等提供了资料。具体做法如下：

(1) 分析砂砾岩油田岩心，确定岩石类型数目，在有岩心的钻孔剖面上采集不同岩性对应的各种测井参数值，作为判别分析的样本。

(2) 根据不同岩石类型的样本值，建立识别岩性的判别函数 $F_i(X)$。

(3) 把具有岩心井段的测井参数曲线离散抽样输入计算机，利用已建立的判别函数对岩心井段的岩性进行识别，以检验判别函数的可靠性。

(4) 把没有取芯的各井测井参数曲线用数字化仪离散取样输入计算机。

(5) 把不同深度点上各测井参数的离散抽样值代入判别函数 $F_i(X)$，以 $F_i(X)$ 为最大对采样点的岩性进行归类，并记录下归类号、相应的深度及测井参数。

(6) 根据上一步记录结果，由计算机绘制岩性剖面及相应的测井参数曲线图。

某砂砾岩油田的岩芯可分出砾岩、砂岩和泥岩三种岩石类型，在岩性剖面上取了岩性可靠

的 84 个样品(其中砾岩样品 30 个、砂岩样品 30 个、泥岩样品 24 个)及其对应的 9 项测井参数，引人 7 项参数，得识别砾、砂和泥岩的判别函数如下：

$$
\begin{aligned}
F_1(X) &= 1.7856x_1 + 0.6465x_3 + 0.1558x_4 + 23.6036x_6 \\
&\quad + 1.7561x_7 + 14.5060x_8 - 0.2027x_9 - 205.3920 \\
F_2(X) &= 1.1269x_1 - 0.4794x_3 - 0.1506x_4 + 16.7496x_6 \\
&\quad + 1.8732x_7 + 14.8695x_8 - 2.4299x_9 - 197.8605 \\
F_3(X) &= -0.0545x_1 + 0.3135x_3 + 0.2032x_4 + 18.9497x_6 \\
&\quad - 2.6158x_7 + 17.5578x_8 - 1.5427x_9 - 287.9910
\end{aligned}
$$

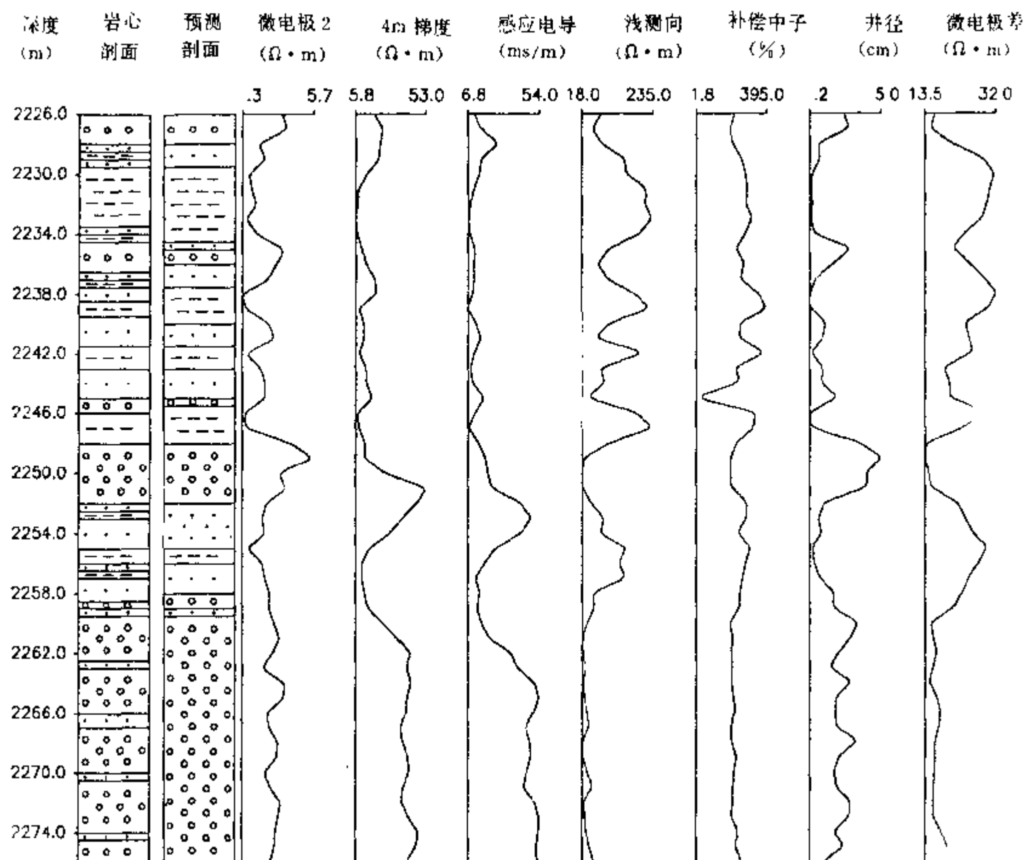


图 5-3 岩性剖面及部分电测曲线图

上述判别函数对 84 个岩石样品中砾岩、砂岩和泥岩的正判率分别为 93%、97%和 96%，平均正判率为 95%，判别验证的正判率从一个方面反映了判别函数对岩性识别的可靠性。但是，这些样品取自不同井的岩心，他们并不代表一个连续的地层剖面。因此，正判率还不能充分说明判别函数对一个地层剖面的识别效果。为此，又利用上述判别函数识别了研究区内永 1-5 井 2226-2276 m 岩心段所对应层段的岩性，以此来检验判别函数对地层剖面进行岩性识别的有效性。岩心剖面与识别的地层岩性剖面——预测剖面如图 5-3 所示。

两个剖面在局部上存在着岩性的差异，造成这一差异的主要原因是利用砾岩、砂岩和泥岩

的代表性样品所确定的判别函数，对过渡性岩层进行识别时，就会出现岩性上的偏差。另外，上、下岩层对薄夹层测井参数的影响和对测井参数的取样间隔也是造成岩性误识的因素。从整体上看，预测剖面与岩心剖面吻合较好，基本上反映了地层的岩性特征。在此基础上，利用上述判别函数识别了砾岩体上各井中地层的岩性。

【例 4】 沉积相定量判别

东濮凹陷西部沙三段有三种沉积相、即三角州、浊流和风暴流。对取自这三种沉积相的 45 个岩样在显微镜下统计了成份成熟度 x_1(石英/(长石＋岩屑))、杂基含量 x_2 和胶结物含量 x_3 三个参数，同时还计算了粒度参数 M_z、σ、S_{k1} 和 k_g，分别建立了沉积相判别函数。

成份参数判别函数：

$$F_1(X) = 2.2841x_1 + 35.3609x_2 + 40.6390x_3 - 10.3050$$
$$F_2(X) = 1.1325x_1 + 45.3407x_2 + 25.9479x_3 - 7.4934$$
$$F_3(X) = 1.6060x_1 + 33.2762x_2 + 36.2213x_3 - 7.9645$$

粒度参数判别函数：

$$F_1(X) = 6.2442M_z + 4.4746\sigma - 4.0192S_{k1} + 11.0180k_g - 24.8472$$
$$F_2(X) = 4.9792M_z + 7.5834\sigma + 0.9221S_{k1} + 8.3104k_g - 22.0130$$
$$F_3(X) = 6.4199M_z + 8.1551\sigma + 1.4336S_{k1} + 9.6124k_g - 31.3128$$

样品参数及判别检验结果见表 5-4。

表 5-4 沉积岩样品参数及判别结果

沉积相号	样品号	成份参数			粒度参数				成份参数后验概率及判别相号	原相号	粒度参数判别相号及后验概率
		x_1	x_2	x_3	M_z	σ	s_{k1}	k_g			
I(三角洲)	1	1.50	0.01	0.33	3.47	0.59	0.39	1.40	0.5314 1	1	1 0.8101
	2	2.67	0.07	0.11	2.85	1.70	0.45	2.04	0.4525 1	1	2 0.6410
	3	1.33	0.30	0.01	4.70	1.56	0.43	1.46	0.9515 2	1	3 0.4367
	4	4.00	0.07	0.08	3.10	2.00	0.45	1.63	0.6572 1	1	2 0.7603
	5	1.50	0.05	0.33	3.55	0.53	0.47	1.55	0.5477 1	1	1 0.8460
	6	2.12	0.06	0.29	3.20	0.99	0.10	1.29	0.6091 1	1	1 0.7403
	7	1.50	0.10	0.19	3.67	1.22	0.41	2.34	0.4689 3	1	1 0.8404
	8	2.50	0.08	0.19	4.13	0.67	0.26	1.91	0.5481 1	1	1 0.9749
	9	2.14	0.06	0.26	4.57	1.37	0.47	1.56	0.5756 1	1	1 0.4286
	10	1.25	0.05	0.35	3.48	1.20	0.30	2.46	0.5285 1	1	1 0.9196
	11	1.22	0.06	0.35	4.54	1.19	0.28	1.73	0.5271 1	1	1 0.8383
	12	6.14	0.10	0.40	3.69	0.89	0.67	2.61	0.9780 1	1	1 0.8795
	13	1.50	0.10	0.22	3.69	0.90	0.51	1.75	0.4670 3	1	1 0.7153
	14	3.00	0.35	0.01	3.34	0.81	0.43	2.61	0.9017 2	1	1 0.9646
	15	4.90	0.30	0.05	3.46	0.87	0.42	2.41	0.5797 1	1	1 0.9427

续表 5-4

沉积相号	样品号	成份参数			粒度参数				成份参数后验概率及判别相号	原相号	粒度参数判别相号及后验概率
		x_1	x_2	x_3	M_z	σ	s_{k1}	k_z			
2（浊流）	16	0.50	0.28	0.01	4.57	1.40	0.35	1.45	0.9639 2	2	1 0.5048
	17	2.33	0.26	0.09	3.21	1.51	0.54	1.73	0.6885 2	2	2 0.6617
	18	1.50	0.25	0.20	2.30	1.78	0.46	1.45	0.4969 2	2	2 0.8611
	19	0.25	0.16	0.06	3.95	1.60	0.35	1.31	0.8143 2	2	2 0.5509
	20	0.82	0.20	0.06	4.00	1.58	0.43	1.42	0.8310 2	2	2 0.5130
	21	1.97	0.20	0.02	3.88	2.21	0.64	1.38	0.7810 2	2	2 0.5749
	22	2.33	0.16	0.00	3.13	1.49	0.60	2.46	0.6891 2	2	1 0.4225
	23	1.97	0.20	0.01	1.57	1.00	0.35	1.20	0.8007 2	2	2 0.9122
	24	1.04	0.10	0.05	1.66	1.67	0.59	0.85	0.6010 2	2	2 0.9881
	25	2.33	0.22	0.07	4.39	1.62	0.56	1.37	0.6470 2	2	3 0.4492
	26	2.41	0.25	0.10	2.62	1.75	0.35	1.34	0.6198 2	2	2 0.8920
	27	1.50	0.18	0.11	4.73	2.02	0.54	1.39	0.5778 2	2	3 0.6532
	28	2.59	0.25	0.06	4.55	1.86	0.61	1.51	0.6995 2	2	3 0.6088
	29	1.00	0.45	0.01	2.55	1.85	0.41	1.82	0.9927 2	2	2 0.8180
	30	1.22	0.14	0.01	3.17	1.24	0.13	1.21	0.7705 2	2	2 0.4882
	31	2.41	0.25	0.04	3.57	1.46	0.62	1.73	0.7766 2	2	2 0.567
	32	1.04	0.20	0.02	4.08	1.53	0.54	1.47	0.8706 2	2	2 0.49
	33	1.67	0.19	0.06	3.23	2.07	0.48	1.36	0.7108 2	2	2 0.7960
3（风暴流）	34	1.00	0.05	0.20	3.72	0.94	0.44	1.87	0.5737 3	3	1 0.8099
	35	2.03	0.05	0.40	4.60	2.24	0.78	2.24	0.7037 1	3	3 0.8540
	36	1.20	0.06	0.10	4.11	1.05	0.44	1.23	0.4871 3	3	1 0.4826
	37	2.00	0.14	0.14	5.54	2.31	0.64	1.34	0.3657 3	3	3 0.8807
	38	1.58	0.13	0.17	5.03	2.76	0.73	1.69	0.4140 3	3	3 0.8892
	39	1.06	0.08	0.27	4.34	1.96	0.57	0.91	0.5186 3	3	2 0.6023
	40	2.33	0.07	0.08	4.57	2.22	0.60	1.32	0.4322 3	3	3 0.6329
	41	1.30	0.15	0.20	4.25	2.21	0.63	1.91	0.4088 3	3	3 0.6895
	42	3.55	0.11	0.11	3.95	1.62	0.61	2.89	0.6159 1	3	1 0.5801
	43	1.50	0.15	0.30	4.45	1.77	0.43	1.94	0.5288 1	3	3 0.4860
	44	1.50	0.19	0.16	3.68	1.61	0.58	1.93	0.4552 2	3	2 0.4538
	45	2.33	0.13	0.12	3.22	1.74	0.27	1.21	0.3740 1	3	2 0.8002

从判别效果来看，成份参数明显优于粒度参数，其原因在于粒度参数的多解性，即沉积物的粒度受沉积环境水动力条件的控制，但不同的沉积环境可具有相似的水动力条件。

【例 5】 利用判别分析预报油气勘探成功率。陈立平，陈子恩等利用判别分析，对四川盆地侏罗系自流井群大安寨组评价区的勘探成功率进行预测。评价区单元划分，变量选取等均与回归中的例 5 相同。675 个单元有钻探资料的 139 个单元中，57 个单元获得了工业油气井，令其勘探成功率为 1，为 A 组，未获得工业油气井、井经研究认为也不可能获得工业油气井的有 27 个单元，令其成功率为 0，为 B 组。以 14 个地质变量对两组作 Bayes 逐步判别，取 $f_1=f_2=2$，得到的两个判别函数如下。

$$F_A = 1.3911 \times 10^{-2}x_2 + 3.83 \times 10^{-3}x_3 + 1.6138 \times 10^{-2}x_4 + 5.27132 \times 10^{-1}x_5 + 8.9093 \times 10^{-2}x_7 - 34.9155$$

$$F_B = 1.7956 \times 10^{-2}x_2 + 7.663 \times 10^{-3}x_3 + 2.1202 \times 10^{-2}x_4 + 2.91067 \times 10^{-1}x_5 + 3.2441 \times 10^{-2}x_7 - 37.3755$$

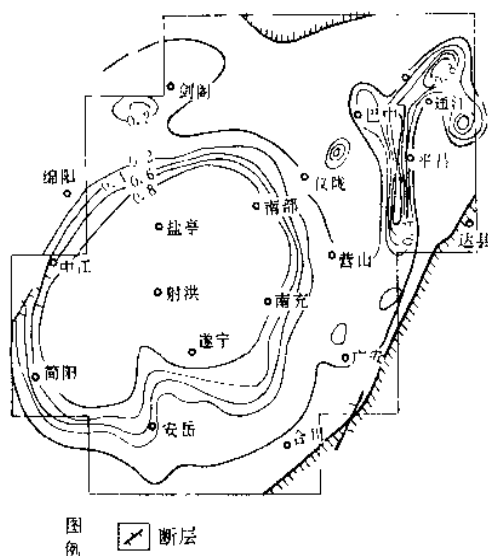


图 5-4 四川盆地侏罗系自流井群大安寨组
勘探成功概率预报图

F_A 是成功率为 1 的 A 组的判别函数，F_B 是成功率为 0 的 B 组的判别函数。选入的变量 x_2、x_3、x_4 属于构造因素，变量 x_5 表征了生储搭配，而 x_7 代表生储条件。这五个变量的组合反映了受岩性岩相控制的裂缝性油藏的特点。

该判别函数对原样品的回判为：原 A 组 57 个中判对 54 个，B 组 38 个判对 36 个，判对率都在 94.7%以上，判别函数是高度显著的。

将评价单元的五个地质变量 x_2、x_3、x_5、x_7 代入判别函数，进一步求出属于 A 组的后验概率，它就是该单元的勘探成功率，在此基础上绘出了勘探成功率预测等值线图，如图 5-4 所示。

习　　题

1. 什么是判别分析？

2. 试述建立线性判别函数的费歇尔准则。

3. 如何利用线性判别函数对样品所属的总体作出判别？

4. 在多总体判别分析中，构造一般判别函数的基本思路是什么？

5. 为何要采用逐步判别？

6. 试述逐步判别的基本思想及计算步骤。

7. 对于例 4 给出的数据，若取 $a=0.1$，试用逐步判别程序建立三角州、浊流、风暴流沉积的判别函数并对样品 $X=(2.67、0.07、0.01)$，$Y=(1.97、0.20、.02)$，$Z=(1.58、0.13、0.107)$ 的沉积相作出判别。

第六章　趋势面分析

地质上的一些变量，如地层界面的埋藏深度、某储集层中油气的比重和流体的压力、某个油气地表化探指标等，都可以看作是分布在空间中的某个曲面 G 上。这样，任何一个地质变量 z 的观测值 z_i 及其观测点的地理坐标 (x_i, y_i) 就构成曲面 G 上的已知点，记为

$$M_i(x_i, y_i, z_i) \quad (i = 1, 2, \cdots, n) \tag{6-1}$$

趋势面分析就是根据 G 上的已知点 $M_i(x_i, y_i, z_i)(i=1,2,\cdots,n)$ 拟合一个数学曲面 L，以此研究地质变量 z 在区域上和局部范围内变化特征的一种统计分析方法。拟合出的数学曲面 L 叫做趋势面，它并不是地质变量 z 分布的实际曲面 G，而是一个逼近于 G 的数学曲面。因此，实测点 $M_i(x_i, y_i, z_i)(i=1,2\cdots,n)$ 分布在趋势面的上、下附近或趋势面上，如图 6-1 所示。

任何一个地质变量 z 的观测值都是由趋势值（背景值）、异常值和随机干扰值组成，即

$$z_i = \hat{z}_i + u_i + v_i \tag{6-2}$$

式中　z_i——地质变量观测值；

$\hat{z}_i$——受区域性地质因素控制的地质变量趋势值；

u_i——局部地质因素决定的地质变量的局部异常值；

v_i——由随机因素造成的干扰值。

在地质研究工作中，特别是与找矿有关的一些地质问题，往往都可归结为寻找区域上的局部异常问题。趋势面分析的目的就是从地质变量 z 的观测值 z_i 中分离出它的趋势部分 $\hat{z}_i$ 和局部异常 u_i，为找矿提供信息。

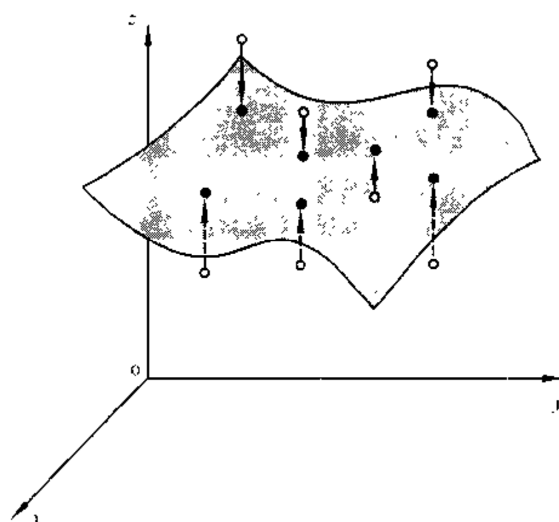


图 6-1　趋势面示意图

●实测点在趋势面上的投影点　○实测点

在趋势面分析中，多项式和傅立叶级数是最常用的 2 种数学模型。

§1　多项式趋势面分析

一、多项式曲面的一般形式

多项式曲面的一般形式如下：

$$z = \beta_1 + \beta_2 x + \beta_3 y + \beta_4 x^2 + \beta_5 xy + \beta_6 y^2 + \cdots \tag{6-3}$$

式中　z——地质变量；

x、y——是地质变量 z 的地理坐标。

若式(6-3)中自变量的最高次数为 k，则称式(6-3)为 k 次多项式曲面。当 $k=1$ 时，它是空

间中的一个平面，如图 6-2(a)所示；当 $k=2$ 时，它为一个抛物面、椭球面或双曲面，如图 6-2(b)所示；随着 k 的增大，曲面的形态就变得更复杂，如图 6-2(c)所示。

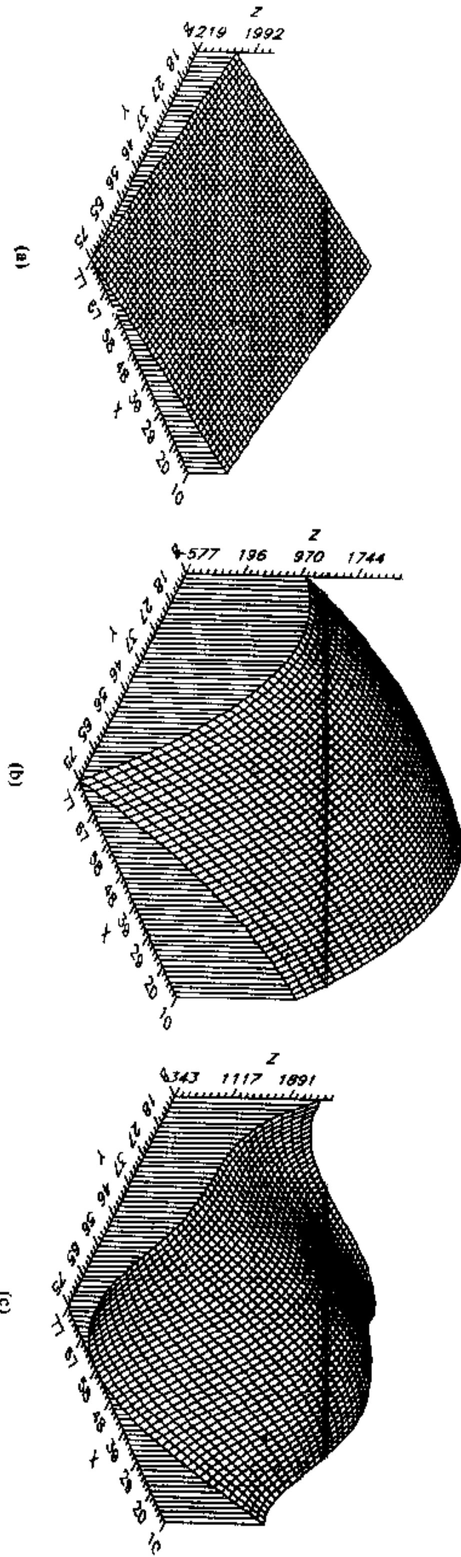


图　　多项式曲面
(a)一次多项式曲面
(b)二次多项式曲面
(c)四次多项式曲面

二、确定多项式的系数

具体地说，要拟合表征地质变量空间分布规律的多项式曲面，就要根据已知点的地理坐标及变量的观测值确定式(6-3)中系数 $\beta_1,\beta_2,\beta_3,\cdots$ 的估计值 $b_1,b_2,b_3,\cdots$。

1. 原始数据及多项式趋势面

设对地质变量 z 进行了 n 次观测，得 n 组观测值如下：

$$(x_i,y_i,z_i)\quad(i=1,2,\cdots,n)$$

并且由 n 组观测值求得 $\beta_1,\beta_2,\beta_3,\cdots$ 的估计值 b_1，$b_2,b_3,\cdots$，得多项式曲面方程

$$\hat{z}=b_1+b_2x+b_3y+b_4x^2+b_5xy+b_6y^2+\cdots \tag{6-4}$$

式(6-4)称为多项式趋势面。

2. 多项式系数的确定

把观测点的地理坐标 $(x_i,y_i)(i=1,2,\cdots,n)$ 代入式(6-4)，得

$$\hat{z}_i=b_1+b_2x_i+b_3y_i-b_4x_i^2+b_5x_iy_i+b_6y_i^2+\cdots$$

确定 $\beta_1,\beta_2,\beta_3,\cdots$ 估计值 $b_1,b_2,b_3,\cdots$ 所依据的原则是要求

$$Q_1(b_1,b_2,b_3,\cdots)=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2$$

达到最小。$Q_1(b_1,b_2,b_3,\cdots)$ 是关于系数 $b_1,b_2,b_3,\cdots$ 的二次函数，根据极值原理，有

$$\begin{cases}\dfrac{\partial Q_1(b_1,b_2,b_3,\cdots)}{\partial b_1}=0\\ \dfrac{\partial Q_1(b_1,b_2,b_3,\cdots)}{\partial b_2}=0\\ \vdots\qquad\qquad\vdots\\ \dfrac{\partial Q_1(b_1,b_2,b_3,\cdots)}{\partial b_p}=0\end{cases} \tag{6-5}$$

式(6-5)中 b_p 为多项式的第 p 个系数。

对于一次多项式，化简整理式(6-5)，可以得到如下正规方程组

$$\begin{cases} b_1\sum 1 + b_2\sum x_i + b_3\sum y_i = \sum z_i \\ b_1\sum x_i + b_2\sum x_i^2 + b_3\sum x_i y_i = \sum z_i x_i \\ b_1\sum y_i + b_2\sum x_{ii} + b_3\sum y_i^2 = \sum z_i y_i \end{cases}$$

写成矩阵形式有：

$$A \cdot B = C \tag{6-6}$$

式中

$$A = \begin{bmatrix} \sum 1 & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i^2 & \sum x_i y_i \\ \sum y_i & \sum x_i y_i & \sum y_i^2 \end{bmatrix} \quad B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad C = \begin{bmatrix} \sum z_i \\ \sum z_i x_i \\ \sum z_i y_i \end{bmatrix}$$

由式(6-6)可解出系数 b_1, b_2, b_3，得一次多项式趋势面方程

$$\hat{z} = b_1 + b_2 x + b_3 y$$

一般，对于高次趋势面的计算，正规方程组的系数矩阵和常数项矩阵为：

$$A = X'X \text{ , } C = X'Z \tag{6-7}$$

式中

$$X = \begin{bmatrix} 1 & x_1 & y_1 & x_1^2 & x_1 y_1 & y_1^2 & x_1^3 & x_1^2 y_1 & x_1 y_1^2 & y_1^3 & \cdots \\ 1 & x_2 & y_2 & x_2^2 & x_2 y_2 & y_2^2 & x_2^3 & x_2^2 y_2 & x_2 y_2^2 & y_2^3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ 1 & x_n & y_n & x_n^2 & x_n y_n & y_n^2 & x_n^3 & x_n^2 y_n & x_n y_n^2 & y_n^3 & \cdots \end{bmatrix}$$

$$Z = (z_1 z_2 \cdots z_n)'$$

三、趋势面的拟合度

所谓趋势面的拟合度，是指观测点上的趋势值与实测值在总体上的逼近程度。

如果记

$$Q = \sum_{i=1}^{n} (z_i - \bar{z})^2$$

$$Q_2(b_1, b_2, \cdots) = \sum_{i=1}^{n} (\hat{z}_i - \bar{z})^2$$

其中 z_i、$\hat{z}_i$ 为第 i 个观测点上地质变量的观测值和趋势值，而 $\bar{z}$ 为地质变量观测值的平均值，那么可以证明

$$\begin{aligned} Q &= \sum_{i=1}^{n} (z_i - \hat{z}_i)^2 + \sum_{i=1}^{n} (\hat{z}_i - \bar{z})^2 \\ &= Q_1(b_1, b_2, \cdots) + Q_2(b_1,_2, \cdots) \end{aligned} \tag{6-8}$$

$Q_1(b_1, b_2, \cdots)$是观测值与趋势值之差的平方和，称为偏差（或残差、剩余）平方和，$Q_1(b_1, b_2, \cdots)$越大，拟合程度越低。$Q_2(b_1, b_2, \cdots)$为趋势值与平均值之差的平方和，$Q_2(b_1, b_2, \cdots)$越大，拟合程度越高。为此定义

$$C = \frac{Q_2(b_1,_2, \cdots)}{Q} \times 100\% = (1 - Q_1(b_1, b_2, \cdots))/Q \times 100\% \tag{6-9}$$

为趋势面的拟合度。

当 $C=100\%$时，表示所有趋势值与观测值完全一致，即全部的观测点都在趋势面上；当 $C=70\%$，表示趋势面反映了观测值的 70%，而另外的 30%作为剩余，在趋势面中没有反映出

来。

对于建立的趋势面是否有意义，从数学上可用下面的统计量进行辅助性检验。

$$F=\frac{Q_2(b_{1,2},\cdots)/(L-1)}{Q_1(b_{1,2},\cdots)/(n-L)}=\frac{(n-L)Q_2(b_{1,2},\cdots)}{(L-1)Q_1(b_{1,2},\cdots)} \tag{6-10}$$

服从 $F(L-1,n-L)$ 分布。式中 L 为多项式趋势面方程系数的个数，对于 k 次多项式，$L=(k+1)(k+2)/2$。n 为观测点数。

对于给定的检验水平 α，由 F 分布表查出临界值 F_α。当 $F>F_\alpha$ 时，则认为趋势面反映变量的变化情况是显著的；当 $F\leqslant F_\alpha$ 时，则认为不显著。

四、趋势面偏差图

趋势面偏差图是以观测值与趋势值之差作为绘图数据而绘制的等值线图。在该图上，偏差大于 0 的区域叫正偏差(或正异常、正剩余)区，而偏差小于 0 的区域称为负偏差(或负异常、负剩余)区。正、负偏差区的地质解释要根据地质变量的地质含义作具体分析。例如，若地质变量是地层面埋藏深度的绝对值，那么在该变量的趋势面偏差图图 6-3 上，正异常区在某种意义上是地层面上放大了的局部洼陷，而负异常区为局部突起。这对研究盆地内油气的分布和查找构造图上可能漏掉的圈闭是非常有用的。

五、影响趋势面分析的主要因素

趋势面分析受多种因素的影响，轻则使趋势面变形，重则使趋势面分析失效。这些因素中较重要的是绘图面积内控制点的数目和控制点的分布。

1. 控制点数

绘图面积内控制点的密度直接影响趋势面和偏差图等值线的分布，这是显然的。从数理统计的角度来看，控制点数的下限不得少于多项式方程中系数的个数，否则将使趋势面分析的结果无效。在进行统计假设检验时，控制点数决定着检验统计量 F 的自由度，而自由度必须足够大，以使 F 检验具有统计学意义。

当多项式系数的数目趋近控制点数时，F 的自由度愈来愈小，只有极大的相关系数才是统计学上显著的，并且假设检验的有效性(即不犯第二类错误的概率)随着样本容量的减小而急剧下降。

在进行趋势面分析时，一般不考虑图幅边界以外的控制点，并且绘图面积往往稍稍超出控制点的实际分布范围，在图幅边界上控制点很少，甚至没有控制点，这种情况应尽量避免。因为在上述情况下，几乎无法控制图幅边界附近趋势面的形态。不论控制区内趋势面的坡度如何，将趋势值外推到图幅边界时均将偏离正确的趋势面上的趋势值，并且趋势面的次数越高，则外推结果的偏离就越大，既使控制点在整个图幅范围内直到图幅边界分布很均匀，仍然不可避免地会出现较小的"边界效应"。由上述可知，最好使控制点的分布面积稍大于绘图面积，这样就在图幅边界的边界效应集中区形成一个"缓冲区"，此区内的控制点对绘图面积内的趋势面形态起着约束和控制作用。

2. 控制点的分布

图幅内控制点的分布也是影响趋势面形态的一个重要因素。图 6-4 是某油层的井位分布图，根据图中不同井的井位坐标及对应的油层顶面海拔高程进行三次多项式趋势面分析，结果如图 6-5 所示，图中(a)、(b)、(c)分别是 50(1～50)口井、28(11-37)口井和 36(4、6、10、11-37、39、47、48)口井的资料绘制的趋势面等值线图。由图 6-5 看出，在同样的图幅内，控制点的分布对趋势面有明显的影响，因此，在对地质变量进行趋势面分析时，应注意观测点的分布状况，当

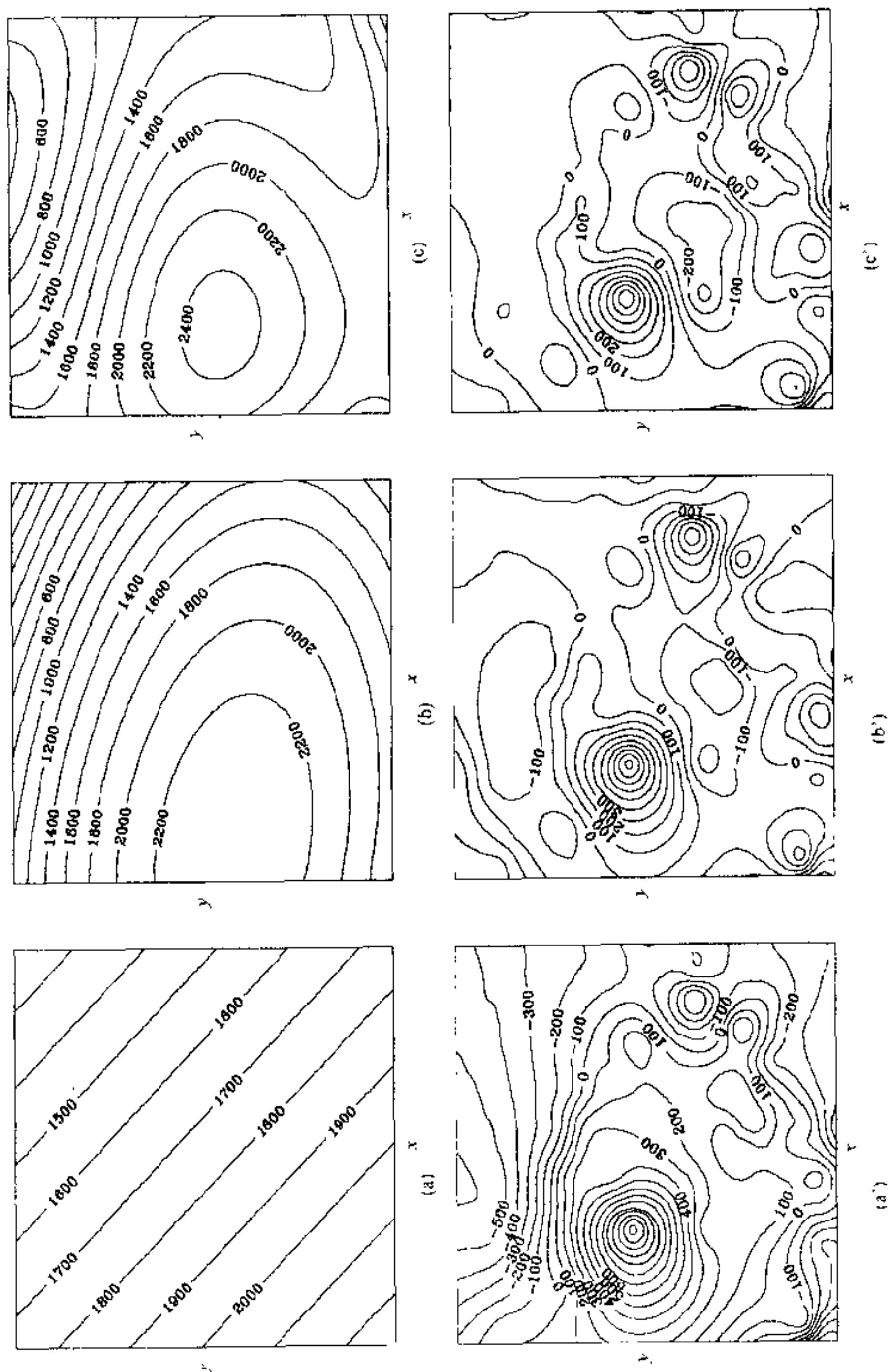


图 6-3 趋势面与偏差等值线图

(a)一次趋势面等值线图 (b)二次趋势面等值线图 (c)四次趋势面等值线图

(a')一次趋势面偏差等值线图 (b')二次趋势面偏差等值线图

(c')四次趋势面偏差等值线图

出现分布不合理时，就要先采取适当的方法加以处理。

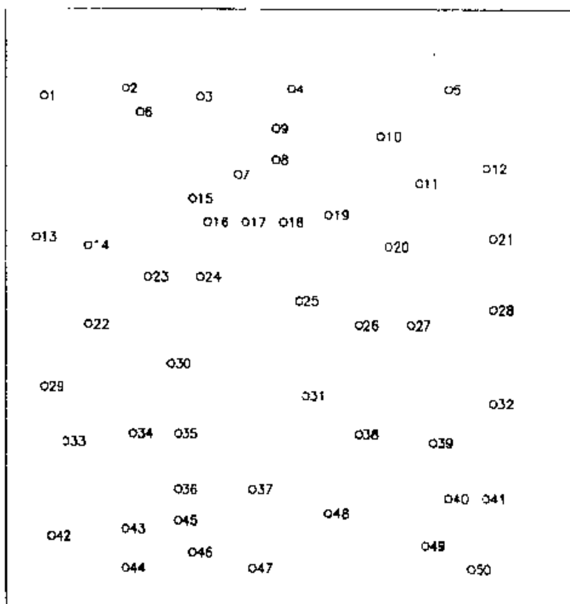


图 6-4 某油层井位分布图

§2 调和趋势面分析

许多地质过程和地质现象具有不严格的周期性特征，如某地层界面的波状起伏、构造旋回，沉积旋回和地球的磁场变化等都表现出周期性的重复现象。对于这种具有周期性的地质过程和现象进行调和趋势面分析，可以把地质变量的区域性趋势部分和剩余部分(周期性部分与随机干扰部分的和)分开，进而研究地质变量的波动特征。

一、正弦波的调和与叠加

1. 正弦波的调和

在不同科学领域内常遇到各种波动现象，其中最简单的波可用正弦波

$$z = a\sin(\omega x + \varphi) \tag{6-11}$$

来描述，其波形如图 6-6 所示。

式(6-11)中的 x 可作为时间变量或空间变量；φ 称为初相位；ω 为角频率；$\lambda=2\pi/\omega$ 为周期或波长；a 称为振幅。式(6-11)还可以表示为初相位为 φ 的正弦波和余弦波的和，即

$$\begin{aligned} z &= a\sin(\omega x + \varphi) = a\sin\varphi\cos\omega x + a\cos\varphi\sin\omega x \\ &= A\cos\omega x + B\sin\omega x \end{aligned} \tag{6-12}$$

式(6-12)中 $A=a\sin\varphi, B=a\cos\varphi$。

反之，任何正弦波和余弦波的和，也都可以表示为单一正弦波的形式。

在此所有由式(6-11)、(6-12)给出的正弦波均称为一维"调和"。当 $\omega=2k\pi/\lambda$ 时，即

$$z_k = a_k\sin\frac{2k\pi}{\lambda}x$$

称为一维 k 阶调和，它的波长为 λ/k。

单个自变量 x 或 y 的正弦(余弦)函数的各种可能的乘积：

$$\begin{matrix} a\cos\omega x\cos v y & b\sin\omega x\cos v y \\ c\cos\omega x\sin v y & d\sin\omega x\sin v y \end{matrix} \tag{6-13}$$

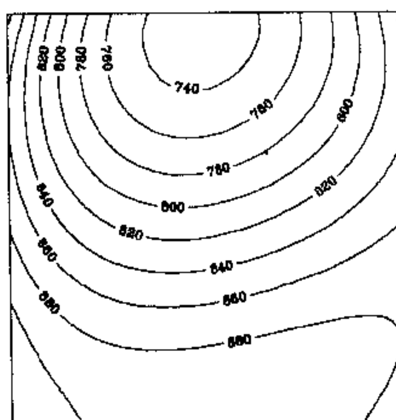
称为二维调和。当式(6-13)中

$$\omega = 2m\pi/\lambda_1 \qquad v = 2n\pi/\lambda_2$$

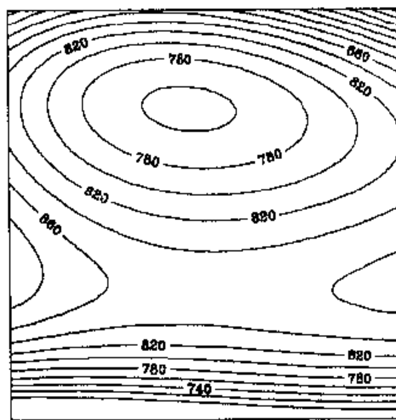
时，称为二维 m、n 阶调和。

2. 正弦波的叠加

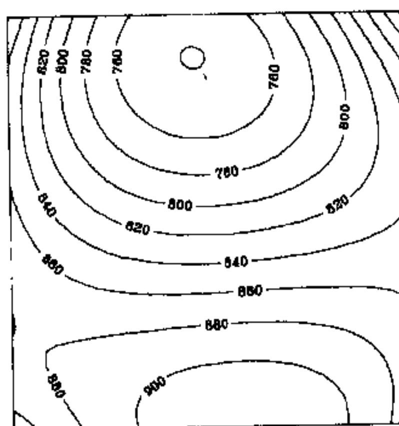
把一些简单的正弦波叠加起来，便可形成各种复杂的波形，如图 6-7 所示，其中(a)、(b)、(c)分别是一阶、二阶、三阶调和的波形，(d)是一阶调和与二阶调和叠加的结果，(e)是一阶调和和三阶调和叠加的结果，(f)是(c)与(d)调和叠加的结果。图中各阶调和的振幅均为 1，实际



(a)



(b)



(c)

图 6-5 油层顶面海拔高程趋势面等值线图

上，振幅可以不相等，因此可以设想，如果利用足够多的具有各种振幅的调和进行叠加，就可以形成各种比较复杂的波形。事实上，当调和的数目无限增大时，就可以综合成非常广泛的一类周期函数；相反，有非常广泛的一类周期函数可以分解成若干(一般为无限个)调和的叠加。

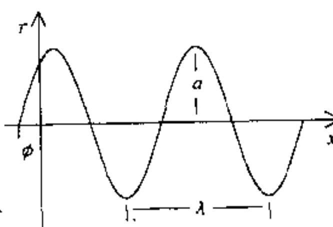


图 6-6　正弦波形示意图

把不同振幅的一维调和进行叠加，可以形成各种比较复杂的曲线。由此可以想象，若把许多简单的二维调和叠加起来，就有可能构造出形态复杂的曲面，如图 6-8 所示。

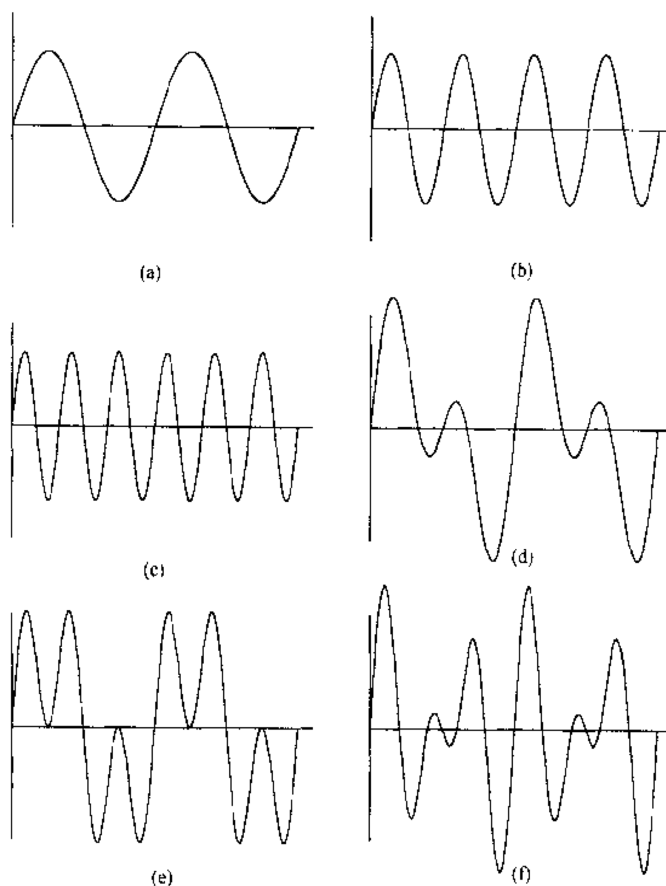


图 6-7　正弦波的调和与叠加

二、调和趋势面

二元傅立叶级数是二维调和的线性组合，它的一般形式为：

$$z = F(x,y) = \sum_{i=0}^{r}\sum_{k=0}^{s}\Big[E_{ik}\cos\frac{2i\pi x}{L}\cos\frac{2k\pi y}{H} + G_{ik}\sin\frac{2i\pi x}{L}\cos\frac{2k\pi y}{H} + P_{ik}\cos\frac{2i\pi x}{L}\sin\frac{2k\pi y}{H} + W_{ik}\sin\frac{2i\pi x}{L}\sin\frac{2k\pi y}{H}\Big] \tag{6-14}$$

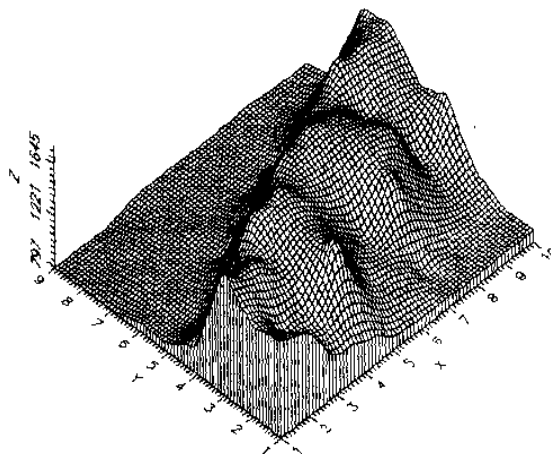


图 6-8　二维调和叠加曲面

1. 调和趋势面方程的建立

建立调和趋势面方程，就是根据曲面上的已知点$(x_i, y_i, z_i)$$(i=1,2,\cdots,n)$求出二元傅立叶级数式(6-14)中$E_{tk}$、$G_{tk}$、$P_{tk}$和$W_{tk}$的估计值$a_{tk}$、$b_{tk}$、$c_{tk}$和$d_{tk}$，得方程

$$\hat{z}=\sum_{t=0}^{r}\sum_{k=1}^{s}\left[a_{tk}\cos\frac{2t\pi x}{L}\cos\frac{2k\pi y}{H}+b_{tk}\sin\frac{2t\pi x}{L}\cos\frac{2k\pi y}{H}\right.$$
$$\left.+c_{tk}\cos\frac{2t\pi x}{L}\sin\frac{2k\pi y}{H}+d_{tk}\sin\frac{2t\pi x}{L}\sin\frac{2k\pi y}{H}\right] \tag{6-15}$$

式中　$\hat{z}$——傅立叶级数趋势值；

r——x方向上给定的傅立叶级数的最大调和阶数；

s——y方向上给定的傅立叶级数的最大调和阶数；

a_{tk}——x方向上为t阶，y方向上为k阶的余弦—余弦项系数；

b_{tk}——x方向上为t阶，y方向上为k阶的正弦—余弦项系数；

c_{tk}——x方向上为t阶，y方向上为k阶的余弦—正弦项系数；

d_{tk}——x方向上为t阶，y方向上为k阶的正弦—正弦项系数；

L——x方向的取样长度，即原图的横向长度；

H——y方向的取样长度，即原图的纵向长度。

为了计算上的方便，傅立叶级数趋势面一般采用其展开形式，例如，一阶(即$r=1$，x方向为一阶；$s=1$，y方向也为一阶)傅立叶级数趋势面方程的展开形式如下：

$$\hat{z}=a_{00}+a_{10}A_1C_0+a_{01}A_0C_1+a_{11}A_1C_1+b_{10}B_1C_0+b_{11}B_1C_1+c_{01}A_0D_1$$
$$+C_{11}A_1D_1+d_{11}B_1D_1 \tag{6-16}$$

一阶傅立叶级数趋势面方程中有9个待定系数：a_{00}，a_{10}，a_{01}，a_{11}，b_{10}，b_{11}，c_{01}，c_{11}，d_{11}。

式中

$$A_t=\cos\frac{2t\pi x}{L},\ B_t=\sin\frac{2t\pi x}{L}\quad(t=0,1)$$

$$C_k=\cos\frac{2k\pi y}{H},\ D_k=\sin\frac{2k\pi y}{H}\quad(k=0,1)$$

为求出方程中的待定系数，可按最小二乘法原理，使每个待定系数对观测值与趋势值的总

离差平方和

$$
\begin{aligned}
Q = \sum_{i=1}^{n}(z_i - \hat{z}_i)^2 = \sum_{i=1}^{n}(z_i - a_{00} - a_{10}A_1C_0 - a_{01}A_0C_1 - a_{11}A_1C_1 \\
- b_{10}B_1C_1 - b_{11}B_1C_1 - c_{01}A_0D_1 - c_{11}A_1D_1 - d_{11}B_1D_1)^2
\end{aligned}
$$

的偏导数等于0,即

$$
\left\{
\begin{aligned}
\frac{\partial Q}{\partial a_{00}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-A_0C_0) = 0 \\
\frac{\partial Q}{\partial a_{10}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-A_1C_0) = 0 \\
\frac{\partial Q}{\partial a_{01}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-A_0C_1) = 0 \\
\frac{\partial Q}{\partial a_{11}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-A_1C_1) = 0 \\
\frac{\partial Q}{\partial a_{10}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-B_1C_0) = 0 \\
\frac{\partial Q}{\partial a_{11}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-B_1C_1) = 0 \\
\frac{\partial Q}{\partial a_{01}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-A_0D_1) = 0 \\
\frac{\partial Q}{\partial a_{11}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-A_1D_1) = 0 \\
\frac{\partial Q}{\partial a_{11}} &= \sum_{i=1}^{n}(z_i - \hat{z}_i)(-B_1D_1) = 0
\end{aligned}
\right.
$$

上述9个方程式经整理可得到一阶傅立叶级数趋势面的正规方程组,可写成如下矩阵形式

$$
\begin{bmatrix}
\sum 1 & \sum A_1C_0 & \sum A_0C_1 & \cdots & \sum B_1D_1 \\
\sum A_1C_0 & \sum (A_1C_0)^2 & \sum A_1C_0A_0C_1 & \cdots & \sum A_1C_0B_1D_1 \\
\sum A_0C_1 & \sum A_0C_1A_1C_0 & \sum (A_0C_1)^2 & \cdots & \sum A_0C_1B_1D_1 \\
\sum A_1C_1 & \sum A_1C_1A_1C_0 & \sum A_1C_1A_0C_1 & \cdots & \sum A_1C_1B_1D_1 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\sum B_1D_1 & \sum B_1D_1A_1C_0 & \sum B_1D_1A_0C_1 & \cdots & \sum (B_1D_1)^2
\end{bmatrix}
\begin{bmatrix}
a_{00} \\ a_{10} \\ a_{01} \\ a_{11} \\ \cdots \\ d_{11}
\end{bmatrix}
=
\begin{bmatrix}
\sum z \\ \sum zA_1C_0 \\ \sum zA_0C_1 \\ \sum zA_1C_1 \\ \cdots \\ \sum zB_1D_1
\end{bmatrix}
\quad (6\text{-}17)
$$

由式(6-17)可解出式(6-16)中的待定系数。

按类似方法,可以得到 x 方向为 r 阶,y 方向为 s 阶的高次调和趋势面的正规方程组。

傅立叶级数趋势面方程的拟合度计算、残差分析方法与多项式趋势面完全相同,在此就不一一叙述了。

2. 调合趋势面方程阶数与待定系数个数之间的关系

由式(6-15)可知调和趋势面方程的值是由傅立叶级数两次求和得到的,即趋势值 $\hat{z}$ 与 x 方向的调和阶数 t 及 y 方向的调和阶数 k 都有关,而求和时的调和阶数都是从0开始的,当 t

$=0, k=0$ 时,因 $\sin 0=0$,所以

$$B_t = \sin \frac{2t\pi x}{L} = 0,\ D_k = \sin \frac{2k\pi y}{H} = 0$$

这就使得 $B_0C_0=A_0D_0=B_0D_0=0$。而 $t=0, k=0$ 时,因 $\cos 0=1$,所以

$$A_t = \cos \frac{2t\pi x}{L} = 1,\ C_k = \cos \frac{2k\pi y}{H} = 1$$

这就使得 $A_0C_0=1$。

可见,在调和趋势面方程中 A_tC_k 项、B_tC_k 或 A_tD_k 项、B_tD_k 项的个数不相同,这就给建立调和趋势面方程的系数矩阵带来一些麻烦。0 至 6 阶调和趋势面方程的 A_tC_k、B_tC_k、A_tD_k、B_tD_k 各项待定系数的个数详见表 6-1。

表 6-1 调和趋势面方程的阶数与待定系数的个数关系表

方程阶数 / 乘积项	0	1	2	3	4	5	6
AC	1(1)	3(4)	5(9)	7(16)	9(25)	11(36)	13(49)
BC	0(0)	2(2)	4(6)	6(12)	8(20)	10(30)	12(42)
AD	0(0)	2(2)	4(6)	6(12)	8(20)	10(30)	12(42)
BD	0(0)	1(1)	3(4)	5(9)	7(16)	9(25)	11(36)
总数	1(1)	8(9)	16(25)	24(49)	32(81)	40(121)	48(169)

表 6-1 中不带括号的数字是对应于方程阶数所新增加的待定系数的个数,而括号中的数字是对应于方程阶数累计待定系数的个数。

调和趋势面方程待定系数的个数由 4 部分组成,即 x 方向为 r 阶,y 方向为 s 阶调和趋势面方程待定系数的个数为

$$\begin{aligned} S &= S_a + S_b + S_c + S_d \\ &= (r+1)(s+1) + r(s+1) + (r+1)s + rs \\ &= 4rs + 2r + 2s + 1 \end{aligned}$$

当 $r=s$ 时,即 x 方向与 y 方向的调和阶数相等时,则有

$$\begin{aligned} S &= S_a + S_b + S_c + S_d \\ &= (r+1)^2 + r(r+1) + (r+1)r + r^2 \\ &= 4r^2 + 4r + 1 \end{aligned}$$

§3 两种模型趋势面分析结果比较

某探区的地质构造具有明显的波状起伏特点。由地震解释资料得到某标准层 90 个观测点的横坐标 x、纵坐标 y 以及层面高程 z,见表 6-2。其中 x、y 是以图幅左下角为坐标原点的相对坐标。采用多项式趋势面分析和调和趋势面分析两种方法进行拟合,目的是把项数相近的两类曲面进行比较。

表 6-2 观测点坐标及标准层高程数据表

序号	横坐标,x	纵坐标,y	高程值,z/m	序号	横坐标,x	纵坐标,y	高程值 z/m
1	1.0	9.0	800.0	46	6.0	5.0	1800.0
2	2.0	9.0	800.0	47	7.0	5.0	1700.0
3	3.0	9.0	800.0	48	8.0	5.0	1150.0
4	4.0	9.0	875.0	49	9.0	5.0	1500.0
5	5.0	9.0	875.0	50	10.0	5.0	1400.0
6	6.0	9.0	875.0	51	1.0	4.0	1040.0
7	7.0	9.0	875.0	52	2.0	4.0	1300.0
8	8.0	9.0	900.0	53	3.0	4.0	1800.0
9	9.0	9.0	875.0	54	4.0	4.0	1800.8
10	10.0	9.0	875.0	55	5.0	4.0	1650.0
11	1.0	8.0	800.0	56	6.0	4.0	1625.0
12	2.0	8.0	800.0	57	7.0	4.0	1650.0
13	3.0	8.0	875.0	58	8.0	4.0	1600.0
14	4.0	8.0	875.0	59	9.0	4.0	1120.0
15	5.0	8.0	875.0	60	10.0	4.0	1475.0
16	6.0	8.0	875.0	61	1.0	3.0	1800.0
17	7.0	8.0	875.0	62	2.0	3.0	1725.0
18	8.0	8.0	875.0	63	3.0	3.0	1550.0
19	9.0	8.0	875.0	64	4.0	3.0	1650.0
20	10.0	8.0	950.0	65	5.0	3.0	1125.0
21	1.0	7.0	800.0	66	6.0	3.0	1200.0
22	2.0	7.0	875.0	67	7.0	3.0	1500.0
23	3.0	7.0	875.0	68	8.0	3.0	1200.0
24	4.0	7.0	900.0	69	9.0	3.0	1000.0
25	5.0	7.0	875.0	70	10.0	3.0	1030.0
26	6.0	7.0	875.0	71	1.0	2.0	1550.0
27	7.0	7.0	950.0	72	2.0	2.0	1300.0
28	8.0	7.0	1100.0	73	3.0	2.0	1300.0
29	9.0	7.0	1330.0	74	4.0	2.0	1700.0
30	10.0	7.0	1850.0	75	5.0	2.0	1200.0
31	1.0	6.0	875.0	76	6.0	2.0	1120.0
32	2.0	6.0	880.0	77	7.0	2.0	930.0
33	3.0	6.0	875.0	78	8.0	2.0	890.0
34	4.0	6.0	890.0	79	9.0	2.0	890.0
35	5.0	6.0	1050.0	80	10.0	2.0	900.0
36	6.0	6.0	1250.0	81	1.0	1.0	1600.0
37	7.0	6.0	1500.0	82	2.0	1.0	1500.0
38	8.0	6.0	1800.0	83	3.0	1.0	1100.0
39	9.0	6.0	1750.0	84	4.0	1.0	1050.0

续表 6-2

序号	横坐标 x	纵坐标 y	高程值 z/m	序号	横坐标 x	纵坐标 y	高程值 z/m
40	10.0	6.0	1700.0	85	5.0	1.0	907.0
41	1.0	5.0	875.0	86	6.0	1.0	890.0
42	2.0	5.0	880.0	87	7.0	1.0	890.0
43	3.0	5.0	1050.0	88	8.0	1.0	890.0
44	4.0	5.0	1325.0	89	9.0	1.0	900.0
45	5.0	5.0	1700.0	90	10.0	1.0	900.0

根据表 6-2 数据，分别进行三次、六次、八次多项式趋势面分析和一阶、二阶、三阶调和趋势面分析，它们的拟合度等参数见表 6-3。三个调和趋势面方程如下：

一阶调和趋势面方程为

$$\hat{z} = 1140.250 - 243.813A_0C_1 - 7.160A_1C_0 + 139.888A_1C_1 + 5.827B_1C_0 + 215.804B_1C_1 + 179.688A_0D_1 - 58.298A_1D_1 - 253.935B_1D_1$$

二阶调和趋势面方程为

$$\begin{aligned}\hat{z} = {} & 1143.158 - 238.638A_0C_1 + 13.727A_0C_2 - 17.388A_1C_0 + 118.150A_1C_1 \\ & - 84.853A_1C_2 + 9.636A_2C_0 + 14.142A_1C_1 + 11.414A_2C_2 - 11.667B_1C_0 \\ & + 180.816B_1C_1 - 122.458B_1C_2 - 8.220B_2C_0 + 79.980B_2C_1 + 10.346B_2C_2 \\ & + 179.489A_0D_1 - 0.939A_0D_2 - 58.695A_1D_1 + 139.234A_1D_2 - 1.587A_2D_1 \\ & + 16.491A_2D_2 + 253.935B_1D_1 - 68.808B_1D_2 - 91.360B_2D_1 - 2.072B_2D_2\end{aligned}$$

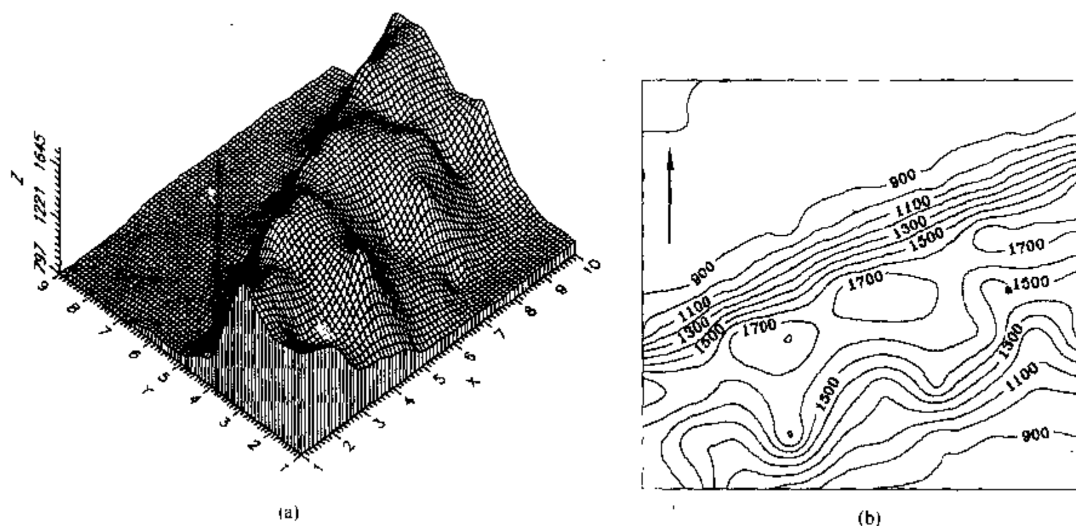
三阶调和趋势面方程为

$$\begin{aligned}\hat{z} = {} & 1157.557 - 224.282A_0C_1 + 24.966A_0C_2 + 47.397A_0C_3 + 15.979A_1C_0 \\ & + 155.921A_1C_1 - 53.316A_1C_2 + 117.409A_1C_3 - 8.181A_2C_0 - 17.731A_2C_1 \\ & - 23.692A_2C_2 - 56.701A_2C_3 + 32.453A_3C_0 - 22.025A_3C_1 - 10.724A_3C_2 \\ & + 13.886A_3C_3 - 13.869B_1C_0 + 176.411B_1C_1 - 126.863B_1C_2 - 11.012B_1C_3 \\ & - 3.388B_2C_0 + 89.645B_2C_1 + 20.010B_2C_2 + 24.162B_2C_3 - 0.876B_3C_0 \\ & + 54.011B_3C_1 - 24.576B_3C_2 - 18.459B_3C_3 + 188.296A_0D_1 - 7.230A_0D_2 \\ & + 5.605A_0D_3 - 41.083A_1D_1 + 126.625A_1D_2 - 0.181A_1D_3 + 16.025A_2D_1 \\ & + 3.909A_2D_2 + 0.945A_2D_3 + 52.837A_3D_1 - 37.745A_3D_2 + 3.185A_3D_3 \\ & + 253.935B_1D_1 - 68.808B_1D_2 + 52.437B_1D_3 + 91.369B_2D_1 - 2.072B_2D_2 \\ & - 12.425B_2D_3 + 62.332B_3D_1 + 20.044B_3D_2 - 36.763B_3D_3\end{aligned}$$

表 6-3 多项式趋势面与二维调和趋势面参数

多项式趋势面				二维调和趋势面			
次数	项数	曲面上的最大极值点数	拟合度/%	阶数	项数	曲面上的最大极值点数	拟合度/%
3	10	1	64.8	1	9	1	73.0
6	28	25	83.8	2	25	16	86.3
8	45	19	86.4	3	49	36	91.3

图 6-9 是表 6-2 数据绘制的曲面图和等值线图，多项式趋势面和调和趋势面图如图 6-10 所示。



(a) (b)

图 6-9 标准层高程曲面图(a)和等值线图(b)

由图 6-9 可见，该标准层具有明显北东－南西延伸的隆起带，其西北侧层面平坦，东南侧层面上具有局部突起和洼陷。

在图 6-10 中，三次多项式趋势面(a)显示出一个大体上不对称的隆起，但是相应的一阶调和趋势面(b)将隆起带勾画得略为清楚一些，这和它们拟合度是相一致的。六次多项式趋势面(c)的拟合度为 83.8%，隆起带的延伸方向更加清楚，并显示出了隆起带上的局部高点，但二阶调和趋势面(d)更好地表现出了层面的波状起伏。与三次多项式趋势面相比，六次多项式趋势面没有明显变化，但三阶调和趋势面却反映出了地层面的基本形态。

由实例比较可知，两种数学模型的拟合效果有差异，那么在实际工作中，究竟选择哪一种拟合方法更好呢？这个问题取决于研究的目的和地质变量的变化特征。由表 6-3 可知，一个包含 45 项的八次多项式能够描述一个有 49 个极值(极大值和极小值)点的曲面，然而一个包含 49 项的三阶二维傅立叶级数却只能表现一个含有 36 个极值点的曲面。由此说明，高次多项式比项数相近的二维傅立叶级数更能反映曲面的复杂性，但是，在反映地质变量周期性变化特征的能力方面，它又不及傅立叶级数。

总的说来，当研究目的是想从观测值中分离出它的周期性部分时，采用调和趋势面分析效果可能更好；如果仅是为了得到简单的或复杂的表现形式，采用多项式趋势面分析方法将更为适合。

§4 多项式趋势面分析源程序

本程序用于求取一批离散数据点的多项式趋势面方程，并计算观测点的趋势值和偏差值；另外，根据需要对趋势值和偏差值进行网格化或多项式曲面插值。现将程序中的主要参数、符号及程序使用方法说明如下：

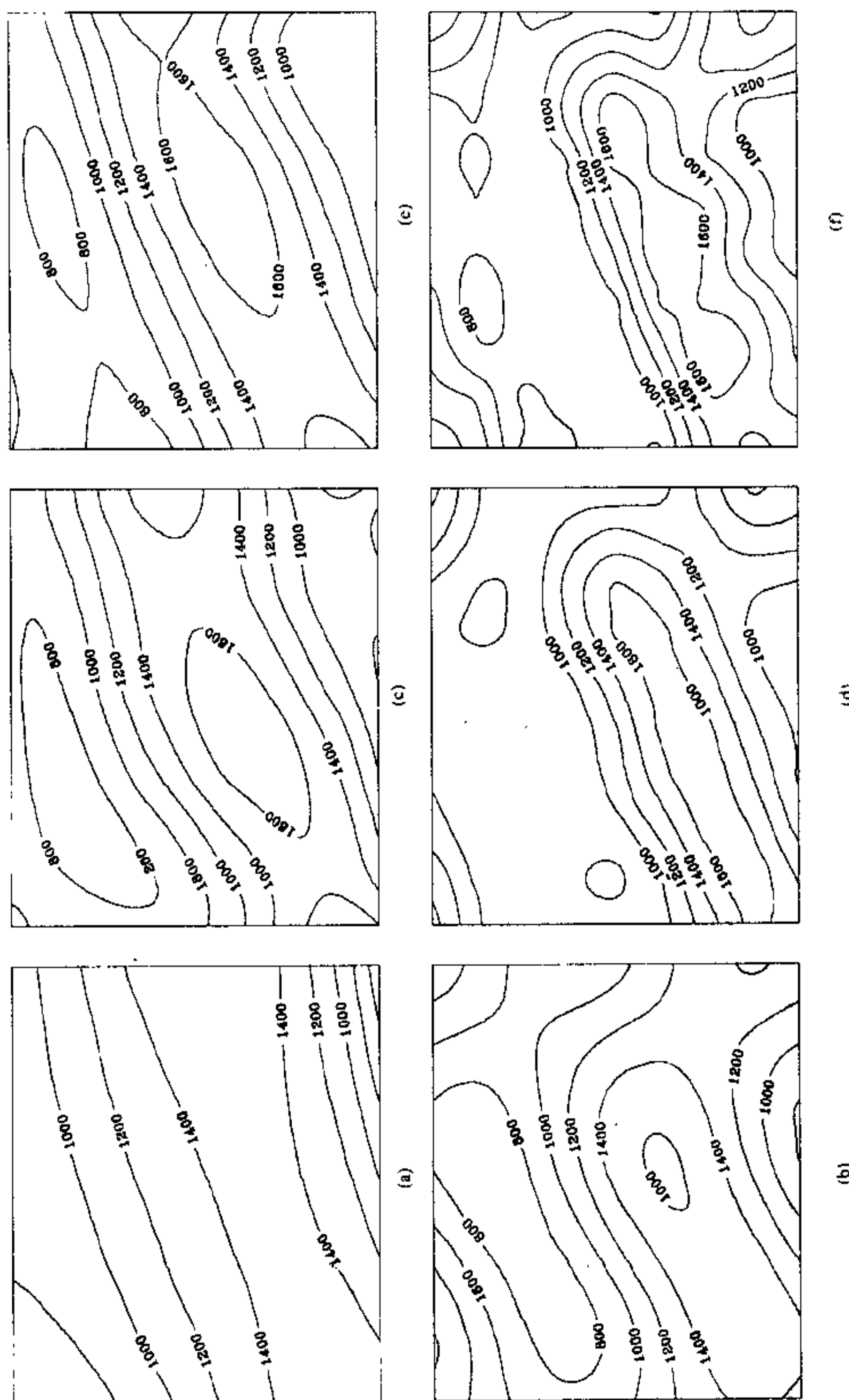


图 6-10　多项式趋势面和二维调和趋势面等值线图

(a)、(c)、(e)为三、六、八次多项式趋势面，(b)、(d)、(f)为一、二、三、阶调合趋势面

一、主要参数及符号

1. 参数

n1——整型变量，观测点的总数；

n2——整型变量，多项趋势面的项数；

Lstart——整型变量，趋势面的最低次数；

Lend——整型变量，趋势面的最高次数；

Lxgrid——整型变量，平行于 x 轴的网格线数；

Lygrid——整型变量，平行于 y 轴的网格线数；

fam——存放原始数据文件名的字符型变量。

2. 符号

x1——存储观测点 x 坐标的一维数组名；

x2——存储观测点 y 坐标的一维数组名；

x3——存储地质变量观测值的一维数组名；

z——存储地质变量趋势值的一维数组名；

rz——存储地质变量偏差值的一维数组名；

yz——存储地质变量网格节点值的二维数组名；

xx——存储网格节点 x 坐标的二维数组名；

yy——存储网格节点 y 坐标的二维数组名；

a——存储正规方程组系数矩阵的二维数组名；

b——存储正规方程组常数项的一维数组名；

ab——存储趋势面方程系数及常数项的一维数组名；

sle——解线性方程组子程序名；

trend——求多项式趋势面方程子程序名；

grid——离散数据网格化子程序名；

insert——多项式曲面插值子程序名。

二、程序运行

1. 数据文件形式

在运行多项式趋势面分析程序之前，首先要建立一个原始数据文件，其形式为

$$(x_i \quad y_i \quad z_i) \qquad (i=1,2,\cdots,n)$$

2. 操作说明

在 DOS 操作系统下键入趋势面分析目标程序名 qsmfx 后，根据屏幕提示做如下操作：

(1) 输入多项式趋势面的最低和最高次数(Input trend surface times: Start to end)：当最低次数和最高次数不相等时，程序只求出从低次到高次多项式趋势面方程，给出相应的拟合度，为选择合适的趋势面提供参考，不作其他运算；当最低次数与最高次数相等时，求出多项式趋势面方程，给出拟合度，并计算观测点的趋势值和偏差值；根据需要还可进行网格化或曲面插值。

(2) 输入数据文件名。

(3) 输入趋势面次数：输入根据拟合度所选择的趋势面次数。

(4) 是否产生网格化数据？若产生键入 Y，否则键入 N。

(5) 在上一步键入 Y 的情况下，接着输入 x 方向和 y 方向的网格线数。

(6) 是否进行曲面插值(Must insert of trend surface: Y/N)

(7)在上一步回答 Y 的情况下,接着键入 x 方向的插值点数(Enter number of points in x dimension) 和 y 方向的插值点数(Enter number of points in y dimension)。

程序运行结束,结果以数据文件形式存盘。

3. 主要输出结果

(1) 趋势面的拟合度;

(2) 各观测点的趋势值和偏差值,其数据文件名分别为 qs.dat 和 rz.dat;

(3) 趋势值和偏差值的网格化数据,数据文件名为 gria.dat;

(4) 多项式曲面插值,数据文件名为 qm.dat。

绘制上述计算结果的曲面图和等值线图的程序见《计算机绘制地质图》一书。

三、源程序

1. 源程序流程

多项式趋势面分析程序流程如图 6-11 所示。

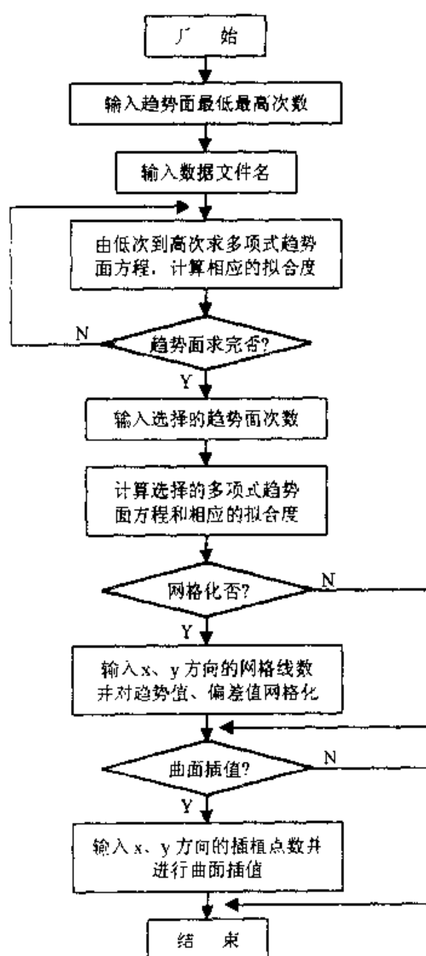


图 6-11 多项式趋势面分析流程图

2. 源程序

```
$large
        parameter(n1=400,n2=45)
        dimension gi(n1,n2),a(n2,n2),b(n2)
        dimension ab(n2),z(n1),rz(n1),yz(20,20)
        common xx(20,20),yy(20,20),zz(20,20)
        common/xyz/ x1(400),x2(400),x3(400)
        common/lg/ lxgrid,lygrid,lxy
        common/Lse/Lstart,Lend
        character *10 fam,fam4,type,type2
        write( * , * ) ' Input trend surface times: start to end '
        read( * , * ) Lstart,Lend
        write( * , * ) 'Input your file name: '
        read( * ,'(a)') fam
        open(1,file=fam)
        nn=0
        do 10 i=1,1000
        read(1, * ,err=20,end=30) x1(i),x2(i),x3(i)
        x1(i)=x1(i) * 0.1
        x2(i)=x2(i) * 0.1
        nn=nn+1
10      continue
        close(1)
20      write( * , * ) ' Err of data file: ',i
        stop
30      write( * , * ) ' End of data file: ',nn
        do 40 i=Lstart,Lend
        m=(i+1) * (i+2)/2
        L=i
        call trend(a,b,ab,gi,z,rz,L,m,nn)
40      continue
        if(Lstart.ne.lend) then
        write( * , * ) ' Input trend surface times: '
        read( * , * ) Lend
        Lstart=Lend
        m=(Lend+1) * (lend+2)/2
        L=Lend
        end if
        call trend(a,b,ab,gi,z,rz,L,m,nn)
        do 41 i=1,10
```

```
41        write( * ,42) ab(i)
42        format(1x,f10.4)
          write( * , * ) ' Creating grid data ? Y/N '
          read( * ,'(a)') type2
          if(type2.eq.'Y'.or.type2.eq.'y') then
          write( * , * ) ' Enter number of grid lines in X dimension: '
          read( * , * ) lxgrid
          write( * , * ) ' Enter number of grid lines in Y dimension: '
          read( * , * ) lygrid
          call grid(nn,1,x1,x2,z)
          do 60 i=1,lxgrid
          do 50 j=1,lygrid
          yz(i,j)=zz(i,j)
50        continue
60        continue
          call grid(nn,1,x1,x2,rz)
          fam4='grid.dat'
          open(4,file=fam4)
          do 80 i=1,lxgrid
          do 70 j=1,lygrid
          write(1,90) xx(i,j),yy(i,j),yz(i,j),zz(i,j)
70        continue
80        continue
90        format(2x,f6.2,',',f6.2,',',f10.2,',',f10.2)
          close(4)
          end if
          write( * , * ) ' Must insert of trend surface ! Y/N '
          read( * ,'(a)') type
          if(type.eq.'Y'.or.type.eq.'y') then
          write( * , * ) ' Enter number of points in X dimension: '
          read( * , * ) Lxgrid
          write( * , * ) ' Enter number of points in Y dimension: '
          read( * , * ) Lygrid
          Lxy=lxgrid * lygrid
          call insert(x1,x2,ab,nn,m,L)
          end if
95        stop
          end

          subroutine trend(a,b,ab,gi,z,rz,L,m,n)
```

```
      dimension gi(n,m),a(m,m),b(m)
      dimension ab(45),z(400),rz(400)
      common/xyz/x1(400),x2(400),x3(400)
      common/lse/Lstart,Lend
      character * 10 fam1,fam2
      do 20 ir=1,n
      do 15 i=0,l
      do 10 j=0,i
      k=i * (i+1)/2+j
      if(x1(ir).eq.0..or.i.eq.j) tHen
      aaa=1
      else
      aaa=x1(ir) * * (i-j)
      end if
      if(x2(ir).eq.0..or.j.eq.0.) tHen
      bbb=1
      else
      bbb=x2(ir) * * j
      end if
      gi(ir,k+1)=aaa * bbb
10    continue
15    continue
20    continue
      do 35 i=1,m
      do 30 j=1,m
      a(i,j)=0.0
      do 25 k=1,n
      a(i,j)=a(i,j)+gi(k,i) * gi(k,j) ! x'x
25    continue
30    continue
35    continue
      do 45 j=1,m
      b(j)=0.
      do 40 i=1,n
      b(j)=b(j)+gi(i,j) * x3(i) ! x'z
40    continue
45    continue
      call sle(a,b,m,ab)
      write( * ,50) l,' tH regression ', 'coefficient '
50    format(1x,i2,a,1x,a)
```

```
      do 60 i=1,n
      z(i)=0.0
      do 55 j=1,m
      z(i)=z(i)+gi(i,j) * ab(j)
55    continue
60    continue
      do 65 i=1,n
65    rz(i)=x3(i)-z(i)
      ss=0.
      do 70 i=1,n
70    ss=ss+rz(i) * * 2
      dcc=float(n)
      ex3=0.0
      do 75 i=1,n
      ex3=ex3+x3(i)
75    continue
      ex3=ex3/dble(dcc)
      s1=0.
      do 80 i=1,n
80    s1=s1+(z(i)-ex3) * * 2
      s=ss+s1
      c=s1/s * 100.0
      write( * ,85) c,s,s1,ss
85    format(1x,' c=',f8.4,/' s=',e12.4,/
     #' s1=',e12.4,/' ss=',e12.4)
      if(Lstart.eq.Lend) then
      fam1=' qs.dat'
      fam2=' rz.dat'
      open(2,file=fam1)
      open(3,file=fam2)
      do 90 i=1,n
      write(2,95) x1(i) * 10,x2(i) * 10,z(i)
      write(3,95) x1(i) * 10,x2(i) * 10,rz(i)
90    continue
95    format(2x,f6.2,',',f6.2,',',f8.2)
      end if
      close(2)
      close(3)
      return
      end
```

```
      subroutine sle(a,b,n,ab)
      dimension a(n,n),b(n),ab(n)
      k1=n-1
      do 70 k=1,k1
      do 10 i=k,n
      if(a(i,k).ne.0.0) go to 20
10    continue
      write(*,21)
21    format(1x,'no unique solution or no solution/')
      stop
20    l=k
      do 30 j=k,n
      if(abs(a(j,k)).gt.abs(a(l,k))) l=j
30    continue
      t=b(k)
      b(k)=b(l)
      b(l)=t
      do 40 j=k,n
      t=a(k,j)
      a(k,j)=a(l,j)
      a(l,j)=t
40    continue
      i1=k+1
      do 60 i= i1,n
      c=a(i,k)/a(k,k)
      b(i)=b(i)-b(k)*c
      do 50 j=i1,n
      a(i,j)=a(i,j)-a(k,j)*c
50    continue
60    continue
70    continue
      ab(n)=b(n)/a(n,n)
      i=n-1
80    ii=i+1
      sum=0.0
      do 90 j=ii,n
90    sum=sum+a(i,j)*ab(j)
      ab(i)=(b(i)-sum)/a(i,i)
      i=i-1
```

```
      if(i.ge.1) go to 80
      return
      end

      subroutine grid(nn,n0,x,y,z)
      dimension x(nn),y(nn),z(nn)
      common xx(20,20),yy(20,20),zz(20,20)
      dimension a1(5),cp(20),b1(20),b2(20),cn(20)
      common/lg/ lxgrid,lygrid,lxy
      call maxmin(x,nn,xmax,xmin)
      call maxmin(y,nn,ymax,ymin)
      dx=(xmax-xmin)/(lxgrid-1)
      dy=(ymax-ymin)/(lygrid-1)
      c=90/n0*0.0174533
      n4=4*n0
      n2=2*n0
      n3=3*n0
      do 10 i=1,n0-1
      a1(i)=tan(c*i)
10    continue
      a1(n0)=9.e+10
      do 30 i=1,lygrid
      yji=ymin+(i-1)*dy
      do 40 j=1,lxgrid
      do 20 ij=1,n4
      b1(ij)=10000.
      b2(ij)=0.
      cp(ij)=1.
20    continue
      xji=xmin+(j-1)*dx
      do 70 k=1,nn
      xk=x(k)-xji
      yk=y(k)-yji
      zk=z(k)
      if(xk.ge.0..and.yk.ge.0.) p=0.
      if(xk.le.0..and.yk.ge.0.) p=n0
      if(xk.le.0..and.yk.le.0.) p=n2
      if(xk.ge.0..and.yk.le.0.) p=n3
      if(xk.eq.0.) u=1.e+6
      if(xk.ne.0..and.n0.eq.1) u=1.
```

```
      if(xk.ne.0..and.n0.ne.1) u=abs(yk/xk)
      v=xk*xk+yk*yk
      do 60 l=1,n0
      if(u.le.a1(l)) then
      l1=l+p
      if(v.le.b1(l1)) then
      b1(l1)=v
      b2(l1)=zk
      end if
      go to 70
      end if
60    continue
70    continue
      pa=0.
      do 80 i1=1,n4
      if(b1(i1).gt.pa) pa=b1(i1)
80    continue
      do 90 i1=1,n4
      do 88 j1=1,n4
      if(j1.ne.i1) then
      cp(i1)=cp(i1)*b1(j1)/pa
      end if
88    continue
90    continue
      w=0.
      do 95 i1=1,n4
95    w=w+cp(i1)
      do 96 i1=1,n4
96    cn(i1)=cp(i1)/w
      zji=0.
      do 50 i1=1,n4
      zji=zji+cn(i1)*b2(i1)
50    continue
      xx(j,i)=xji*10
      yy(j,i)=yji*10
      zz(j,i)=zji
40    continue
30    continue
      return
      end
```

```
      subroutine insert(xx,yy,ab,nn,m,L)
      dimension xx(nn),yy(nn)
      dimension ab(45),gi(400,45)
      dimension xyz(400,3)
      common/lg/lxgrid,lygrid,lxy
      character * 10 fam3
      call maxmin(xx,nn,xmax,xmin)
      call maxmin(yy,nn,ymax,ymin)
      dx=(xmax-xmin)/(lxgrid-1)
      dy=(ymax-ymin)/(lygrid-1)
      fam3='qm.dat'
      open(5,file=fam3)
      xi=xmin-dx
      do 20 i=1,lxgrid
      xi=xi+dx
      i1=(i-1) * lygrid+1
      i2=i * lygrid
      do 15 ii=i1,i2
      xyz(ii,1)=xi
15    continue
20    continue
      do 40 j=1,lygrid
      yi=ymin-dy
      j1=(j-1) * lygrid+1
      j2=j * lygrid
      do 30 jj=j1,j2
      yi=yi+dy
      xyz(jj,2)=yi
30    continue
40    continue
      do 70 ir=1,lxy
      do 60 i=0,L
      do 50 j=0,i
      k=i * (i+1)/2+j
      if(xyz(ir,1).eq.0..or.i.eq.j) then
      aaa=1
      else
      aaa=xyz(ir,1) * * (i-j)
      end if
```

```
      if(xyz(ir,2).eq.0..or.j.eq.0) then
      bbb=1
      else
      bbb=xyz(ir,2)**j
      end if
      gi(ir,k+1)=aaa*bbb
50    continue
60    continue
70    continue
      do 90 i=1,lxy
      xyz(i,3)=0.
      do 80 j=1,m
      xyz(i,3)=xyz(i,3)+gi(i,j)*ab(j)
80    continue
90    continue
      do 95 i=1,lxy
      xyz(i,1)=xyz(i,1)*10
      xyz(i,2)=xyz(i,2)*10
      xyz(i,3)=xyz(i,3)
      write(5,96) xyz(i,1),xyz(i,2),xyz(i,3)
95    continue
96    format(1x,f6.2,',',f6.2,',',f15.4)
      close(5)
      return
      end

      subroutine maxmin(x,n,xmax,xmin)
      dimension x(n)
      xmax=x(1)
      xmin=x(1)
      do 10 i=2,n
      if(x(i).gt.xmax) xmax=x(i)
      if(x(i).le.xmin) xmin=x(i)
10    continue
      end
```

§5 应用算例

【例1】 寻找有利油气储集构造

油气田的勘探经验表明，受构造控制的油气藏占有很大比重，采用传统的地质方法研究构

造与油气关系时，有些局部构造常常被区域构造的展布特性所掩盖，不易很快发现。但趋势面分析却能弥补这方面的缺陷，在分离区域构造背景之后，突出局部构造，为寻找油气田提供新的依据。例如美国勘萨斯州东部的密西西比砾岩，其构造为一区域性的向西倾斜的单斜，其上最大的局部圈闭才 20 呎高，似乎不会形成大的油气藏。但是用趋势面分析之后，存在着大面积

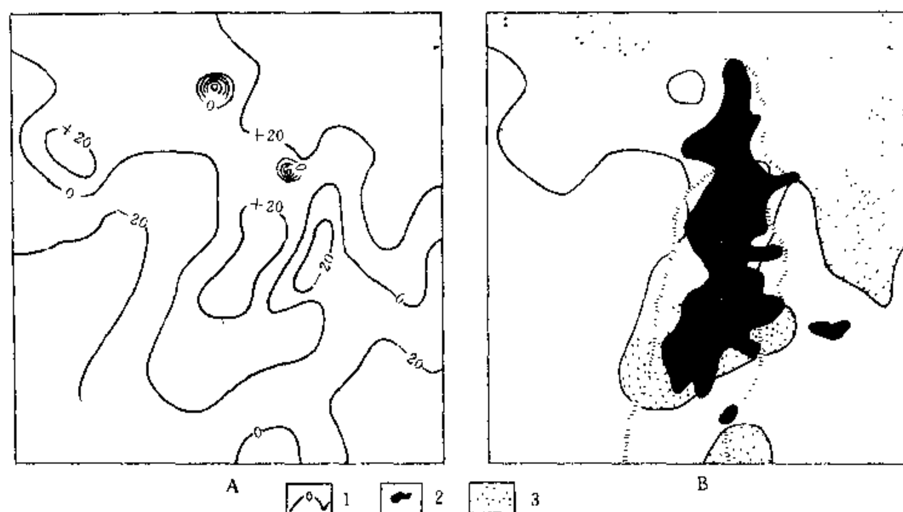


图 6-12

A—勘萨斯中东部密西西比燧石砾岩顶面构造二次趋势面剩余图；B——洛斯特普林油田（黑色）

1. 二次趋势面剩余等值线；2. 1955 年产油区；3. 正剩余区

（据（Merriam 和 Harbaugh，1964）

的正剩余区，与探明的油田分布范围颇为吻合（图 6-12）。图 6-13 是另一个用趋势面分析研究正剩余区与油气田分布关系的例子。从图中可看出大多数油气田均位于正剩余区内。虽然油气的聚集与多种因素有关，尽管如此，趋势面分析仍可为我们提供这一方面的依据，指出构造油气藏的可能分布地区，便于进行勘探部署。值得注意的是，不能单从趋势面分析就得出结论，而应该综合各种地质资料全面解释。既要重视正剩余区的有利部位，也不能忽视负剩余区的有利地带，因岩性油藏和地层油藏可在负剩余区找到。

【例 2】 寻找岩性—构造油气藏

酒西盆地北部单斜带，经过多年勘探在火烧沟群先后发现了白杨河、单北和白东三个油田。能否利用趋势面分析在该区寻找新的岩性—构造油气藏呢？我们和玉门石油研究院勘探室的同志们一起进行了探索，取得了初步效果。

该区第三系火烧沟群构造为一由北向南西倾斜的平缓单斜，在单斜上有少数几个鼻状构造和膝状挠曲。根据趋势面分析，该群顶部构造的背景为一向南西 12°倾斜的平面，倾角为 11°6′。从构造的剩余值图可明显看出（图 6-14），已探明的油田全部位于正剩余区，说明了构造因素在控制油气上占有重要地位。但是白杨河与白东油田并不是在正剩余区的最高部位，而是在斜坡上或靠近正剩余边界的地方。这说明它们除受构造因素控制外，同时更受岩性变化的控制（油层上倾方向物性变差），是属于典型的岩性-构造油藏。研究的结论认为：已知油田都分布在正剩余区地层上倾方向有低渗透带存在的有利构造部位。基于以上认识，提出了几块有利面积（见图 6-14）。经初步钻探，在 A、B、E 三块面积内发现了好的油砂或工业性油流。

【例 3】 研究地下断裂分布

图 6-13 勘萨斯州东南地区(地区Ⅱ)的构造趋势分析结果

A—剩余值图;B—油气田分布图 1. 二次趋势面剩余等值线;

2. 鞋带状砂岩油气田(据 Merriam 和 Harbaugh,1964)

图 6-14 酒西盆地北部单斜带第三系火烧沟群顶一次趋势剩余等值图和有利地区预测图

1—正负剩余区边界;2—剩余值等值线;3—老油田位置;4—预测有利面积

(据陈立官主编《油气田地下地质学》,1983)

趋势面分析的基本功能在于将数据中的区域性背景与局部特征和随机干扰分离出来,以

期从中找出隐蔽的或被掩盖而有意义的信息。在一般情况下，岩体顶面的起伏虽然有较大的随机性，但往往是连续变化。然而构造断裂所造成的起伏则有线状展布、方向性明显和非连续性的特点。这就是地下断裂能通过趋势面分析加以显示的前提。

由于趋势图能表现大范围的总体变化，而剩余图则包含了小范围的局部特征及无规律的随机部份。因此，对剩余值再进行分解，又可得到次一级的趋势图和剩余图。因第一次趋势图已将大范围的变化特征分离出去，于是第二次计算的数据基本上不包含主要的区域性分量了，尤其是在拟合度较高的情况下更是如此。因此，第二次的趋势图表现了小范围的变化特征。而第二次剩余图则更集中地反映了更小的局部特征和随机成份。呈线状延伸，且有一定方向的断裂，必然要在第二次趋势面分析的剩余图中反映出来。

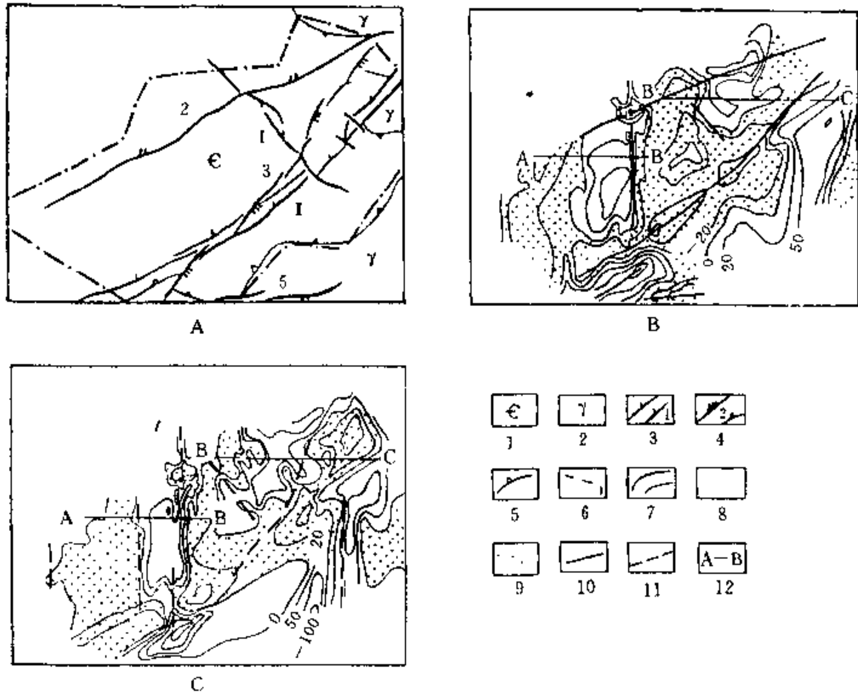


图 6-15　A—某研究区地质图；B—某研究区趋势面分析五次剩余图；C—某研究区第二次的七次剩余图。

1—寒武系地层；2—花岗岩；3—张性、张扭性断裂及编号；4—压性、压扭性断裂及编号；

5—岩体与围岩接触界线；6—数据点分布范围；7—剩余值等值线；8—正剩余区；9—负剩余区

10—已知断裂；11—趋势面分析后预测的断裂；12—剖面线位置及编号

（据陈立官主编《油气田地下地质学》，1983）

图 4-15A 为某地区寒武系地层面资料所绘制的断裂分布图。图 6-15B 为该区五次趋势面分析的剩余图，它反映了地下断裂的情况。从两张图上可明显看出：区内两条 NEE 走向的大断裂是非常吻合一致的。但图 6-15B 中最醒目之处，是中央的 SN 向正剩余区等值线梯度变化特大，使人毫不怀疑这里存在着向东倾斜的断裂，但地面地质图上却无反映。后来通过钻探证明，该区深部确实存在一条较大的断裂。最富有说服力和令人感兴趣的是图 6-15C。它是对第一次五次剩余值进行再次趋势面分析的七次剩余图。可以看出它与图 6-15B 十分相似。从已知断裂部位等值线的方向性、梯度变化、正负线残差值的界线看，都非常吻合，甚至有些部位，把呈平行带状出现的断裂更准确、更细致的反映出来了。从此例不难看出趋势面对断裂构造的

研究十分有用。一般情况下，剩余图中等值线排列的方向性、规模大小和梯度变化，能够反映地下断裂及其产状，尤其对具有一定规模和垂直位移较大的断裂，效果更好。

据胜利油田研究院研究，对于断块油田，利用地下标准层的标高作趋势面分析，在剩余图上等值线密集地方，往往就是断层分布的位置，密集等值线的方向，也就是断层延伸的方向。

【例 4】 预测有利含油区

利用地震波通过含油气层时高频成分被强烈吸收，而低频能量相应增强的特征反演预测油气田的方法称为 HCI 技术，即碳氢检测技术。这种技术可以提供多种资料，地震特征图和地震综合信息图是其中的两种，如图 6-16 和 6-17 所示。

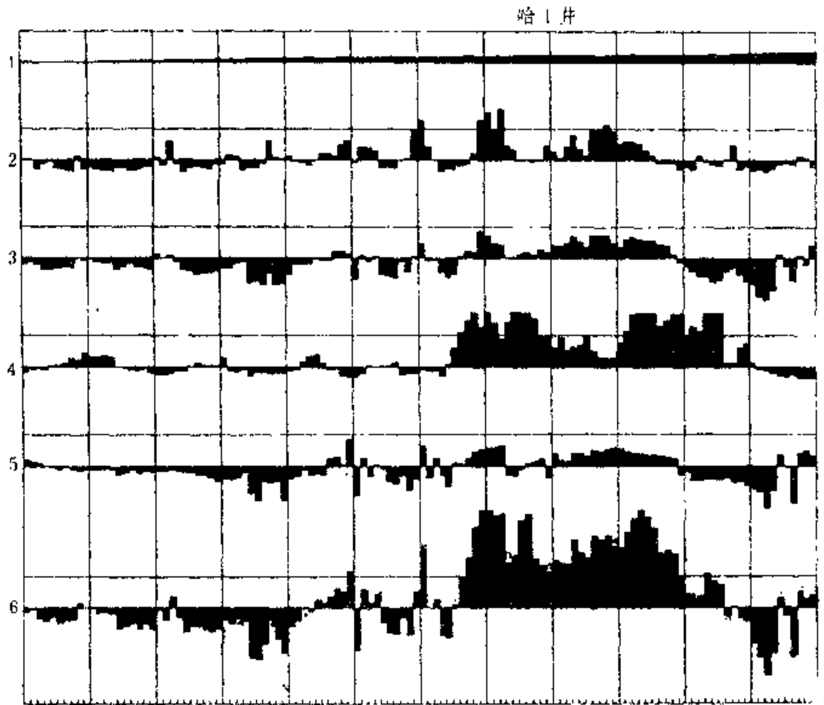


图 6-16 碳氢检测地震特征图
（据张守本等，1984）

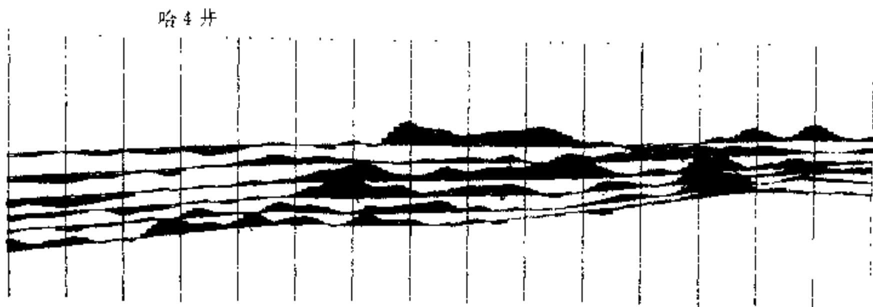


图 6-17 碳氢检测综合图
（据张守本等，1984）

地震特征图由速度变化特征、10赫兹带宽能量百分比、10赫兹能量、平均频率特征、带宽能量特征(谐能)及峰值频率特征等六种地震信息组成。大量资料表明,有利含油气层段在地震特征图上明显地反映出10赫兹能量百分比、10赫兹能量、带宽能量等指标(图6-16中的2、3、5)能量增强,速度、平均频率及峰值频率(图6-16中的1、4、6)降低,而总值升高。

对勘探目的层及其顶和底进行HCI处理,获得诸层各自地震特征图后,再将各层的总值绘制在一张图上,即形成地震信息综合图。

1982年,石油地球物理勘探局第四勘探公司在哈南地区选用了四条HCI剖面,对400个观测点的地震综合信息进行计算,反映上下中三层目的层检测效果。图6-18是哈南地区HCI三次趋势面偏差的四次趋势面图。图中明显地分成两个部分:以斜线阴影表示的负值区和等

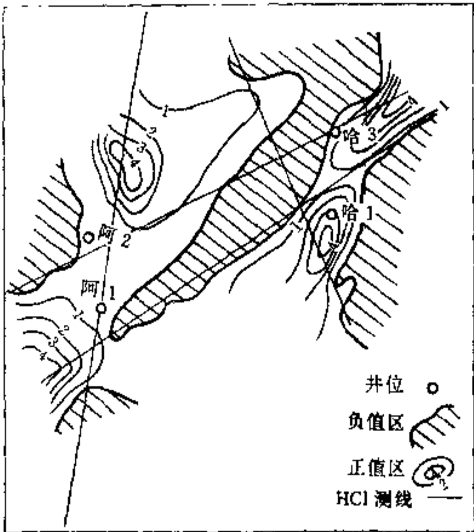


图6-18 哈南地区HCI3 4次偏差分析图
(据张守本,1984)

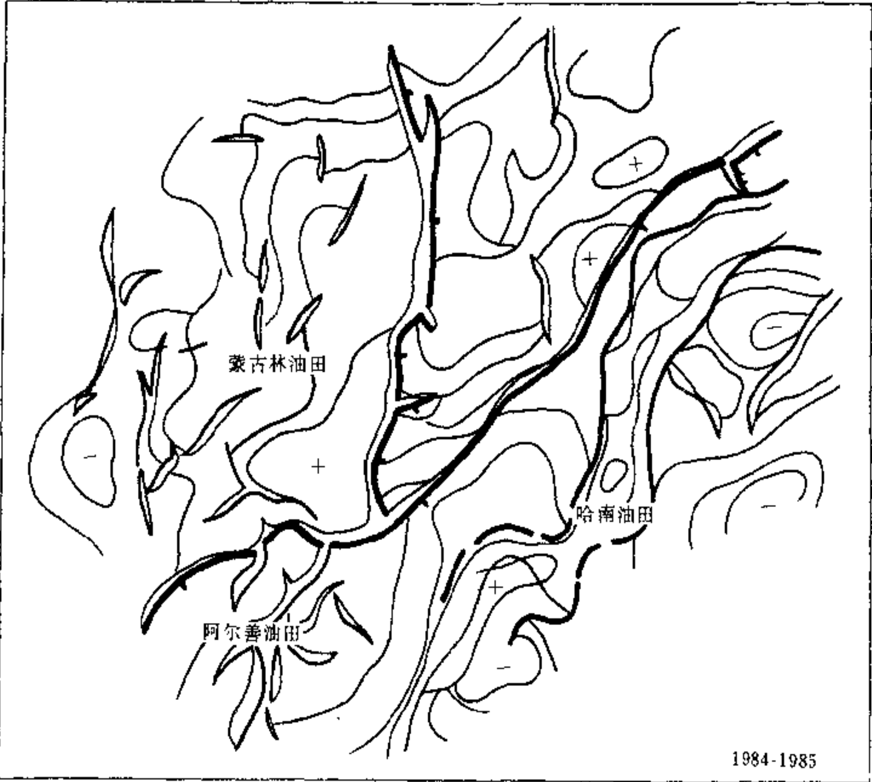


图6-19 HCI后效验证图
(据张守本等,1981)

值线圈闭表示的正值区。正值区为有利的含油气区。在研究区内预测了两个有利地区。东区呈 NNE 走向，是个由两个局部正异常组成的异常带。西区整体亦呈 NE 向，面积为东区的 2 倍，主体在北部，是一个圈闭主体向北开口的正异常区。1984 年以来在相应地区进行地震详查及钻探，发现了哈南油田和蒙古林油田，见图 6-19。原预测的东区发现了一个 NE 走向的潜山背斜构造带。该构造带面积为 $70km^2$，由夫特西、哈南、布敦北和额汉南四个构造组成。在哈南构造上的哈 1 井、哈 8 井均已见自喷油流。原预测的西区在其主体部位发现了一个多高点的背斜构造，经地震和钻探证实是一个具有较大面积、多套储层、较多地质储量的复式油藏。

习　题

1. 什么是趋势面分析？它在石油及天然气地质中有哪些应用？
2. 何谓多项式趋势面分析？何谓调和趋势面分析？各有什么优点？
3. 何谓趋势面的拟合度？它的内涵是什么？是否拟合度越高越好？为什么？
4. 影响趋势面拟合度的主要因素有哪些？
5. 试用矩阵形式写出求二维三次多项式趋势面系数的方程组。
6. 趋势面图和偏差图各具有什么意义？
7. 试比较多项式趋势面分析与多元回归分析在方法原理及应用上的异同。
8. 试举例说明趋势面分析在石油及天然气地质中的应用。

*第七章　因子分析

§1　因子分析概述

一、因子分析的基本概念

因子分析是研究变量间相关关系、样品间相似关系、变量与样品间成因联系以及探索它们之间产生上述关系之内在原因的一些多元统计分析方法的总称。根据它们的研究对象，因子分析大致可分为 R 型因子分析、Q 型因子分析和对应分析三种类型。在此仅介绍基于相关系数和相似系数统计量下的因子分析。

设有 n 个样品，每个样品包含 m 个变量。把 n 个样品 m 个变量的观测值写成数据矩阵形式：

$$Z = \begin{pmatrix} z_{11} & z_{12} & \cdots\cdots & z_{1n} \\ z_{21} & z_{22} & \cdots\cdots & z_{2n} \\ z_{m1} & z_{m2} & \cdots\cdots & z_{mn} \end{pmatrix}$$

在矩阵 Z 中，$z_j=(z_{1j}z_{2j}\cdots\cdots z_{mj})'$ $(j=1,2,\cdots,n)$ 表示第 j 个样品 m 个变量的观测值，即 Z 的第 j 列表示第 j 个样品；$z_i=(z_{i1}z_{i2}\cdots z_{in})$ $(i=1,2,\cdots,m)$ 表示第 i 个变量在 n 个样品中的观测值，即 Z 的第 i 行表示第 i 个变量。

设变量 z_i 的标准差标准化（简称标准化）变量为 x_i，并把相应的观测值记为矩阵

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1n} \\ x_{21} & x_{22} & \cdots\cdots & x_{2n} \\ x_{m1} & x_{m2} & \cdots\cdots & x_{mn} \end{pmatrix}$$

在因子分析中的后续讨论中，若无特殊说明，均约定变量是标准化的。

对于任意两个变量 $x_i=(x_{i1}x_{i2}\cdots\cdots x_{in})$ 和 $x_j=(x_{j1}x_{j2}\cdots x_{jn})$，它们的相关系数为

$$r_{ij} = \frac{1}{n-1}\sum_{k=1}^{n} x_{ik}\cdot x_{jk} \quad (i,j=1,2,\cdots,m)$$

m 个变量的相关系数构成一个 $m\times m$ 的矩阵

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots\cdots & r_{1m} \\ r_{21} & r_{22} & \cdots\cdots & r_{2m} \\ r_{m1} & r_{m2} & \cdots\cdots & r_{mm} \end{pmatrix}$$

且 $r_{ij}=r_{ji}$，$r_{11}=r_{22}=\cdots=r_{mm}=1$，并称 R 为变量的相关矩阵。

任意两个样品 $x_i=(x_{1i}x_{2i}\cdots x_{mi})'$ 和 $x_j=(x_{1j}x_{2j}\cdots x_{mj})'$ 的相似系数为

$$q_{ij} = \frac{\sum_{k=1}^{m} x_{ki}\cdot x_{kj}}{\sqrt{\sum_{k=1}^{m} x_{ki}^2 \sum_{k=1}^{m} x_{kj}^2}} \qquad (i,j=1,2,\cdots,n)$$

n 个样品的相似系数写成矩阵形式为

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots\cdots & q_{1n} \\ q_{21} & q_{22} & \cdots\cdots & q_{2n} \\ q_{n1} & q_{n2} & \cdots\cdots & q_{nn} \end{pmatrix}$$

且 $q_{11}=q_{22}=\cdots=q_{nn}=1$，$q_{ij}=q_{ji}$，称 Q 为样品的相似矩阵。

R 型因子分析是研究相关矩阵 R 的内部结构，从中找出 p 个对所有变量起控制作用的综合变量 $f_k(k=1,2,\cdots,p,p<m)$，并把变量 x_i 表示为 f_k 的线性组合，即

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{ip}f_p + a_ie_i \quad (i = 1,2,\cdots,m) \tag{7-1}$$

当 $p<<m$ 时，由式(7-1)化简研究系统，并进一步探索变量的成因联系及空间中变化规律的控制因素。

Q 型因子分析是通过对相似矩阵 Q 内部结构的研究，寻找制约样品相似性的 p 个综合变量 $f_k(k=1,2,\cdots,p,p<n)$，并把样品 x_j 表示为综合变量 f_k 的线性组合，即

$$x_j = a_{j1}f_1 + a_{j2}f_2 + \cdots + a_{jp}f_p + a_je_j \quad (j = 1,2,\cdots,n) \tag{7-2}$$

当 $p<<n$ 时，由式(7-2)化简样品研究系统，并进一步研究样品产生相似性的主要原因。

在此需要指出的是：式(7-1)和式(7-2)也是 R 型和 Q 型因子分析的基本假设条件。其中 e_i 和 e_j 分别是服从均值为 0、方差为 σ_i^2 和 σ_j^2 的正态分布。

上述两种因子分析方法分析、综合地质数据、揭示变量之间和样品之间在成因上或空间上的联系，但却都不能直接反映出变量与样品间的联系。事实上，在地质学中研究变量和样品之间的联系要比研究它们各自之间的关系更为重要，因为变量与样品之间的联系可以直接用于地质解释。对应分析方法是在上述两种因子分析方法的基础上发展起来的一种多元统计分析方法。它在同一个因子空间中揭示变量与样品的联系，且有 R 型因子分析和 Q 型因子分析的共同特点。

二、因子分析在地质研究中的作用

在目前的观测手段下，地质人员可以获得大量的、特别是数值型的地质资料，这就为地质人员进行地质问题的综合研究创造了有利的条件。但是，地质人员又为如何处理和分析浩瀚而又杂乱，有时又是相互矛盾的地质数据而深感烦恼。因子分析方法在处理和分析具有错综复杂关系的地质数据、进行地质成因推理中可以起到以下诸方面的作用。

1. 压缩原始数据

在收集地质资料时，地质人员总希望收集到尽可能多的地质数据，然而在进行综合研究时却又希望用尽可能少的地质数据阐明所研究的问题。由式(7-1)和式(7-2)可知，因子分析为解决这一问题提供了一条科学的途径。它使原始数据在数量上极大的精简，但并不影响研究结论的可靠性。原因在于 p 个综合变量是对 m 个原始变量起控制作用的主要因素。因此，从地质成因上讲，虽然数据的数量减少了，但却提高了用因于分析方法压缩后的数据在地质成因方面的质量，仍然包含着原始数据所提供的绝大部分成因信息。

由上述可知，因子分析可以起到不损失或少损失地质成因信息、降维分析问题的作用。

2. 指示地质成因的推理方向

因子分析方法能够把庞杂的原始地质数据按照它们在地质成因上的联系进行归纳、整理、精炼乃至分类，从而可以理出几条比较客观的成因线索，这就为人们指出了逻辑推理的方向。

3. 分解叠加的地质过程

任何地质现象一般都可视为多种地质过程叠加的综合产物，既有时间域上不同地质过程的叠加，又有空间域上不同地质过程的影响。这些地质过程相互干扰，把每一个独立地质过程的特征掩盖的面目不清，由此造成地质成因研究的复杂化和多解性。

众所周知，尽管所观察到的地质特征很多，但产生这些地质特征的地质过程却是极为有限的。例如，Cr、Ni、Co、Pt、MgO、FeO、橄榄石、斜方辉石、尖晶石、高的重力值等都可认为是超基性岩浆活动的产物；H_2O、MgO、蛇纹石、岩体内的断裂、岩体同围岩的接触等均同超基性岩的蛇纹石化作用有着密切的联系；而 CaO、CO_2 则同超基性岩的碳酸盐作用密切相关。上述事实表明，各种不同的地质特征都可归结为产生这些地质特征的地质过程。因子分析正是按照地质过程来综合原始数据的统计分析方法。它给出的每一个综合变量代表了一组在成因上密切相关的地质特征，因此，因子分析可以根据地质特征分解叠加的地质过程。

以上三个方面仅仅展示了因子分析目前在地质成因研究中所起的基本作用，至于它在地质学其他方面研究中的潜在能力，尚需继续探索和不断开拓。

三、因子分析在地质研究中的应用

同一种因子分析方法可以用来研究和解决地质学领域内的不同问题。例如，沉积学家可用它来研究沉积环境，矿床学家又可用它来探讨成矿条件，而石油地质学家还可用它来分析盆地的生油条件，因此，不易也不可能给因子分析在地质学中的应用划定一个应用范畴，总的说来，因子分析在各地质分支学科中都获得了广泛的应用，并在以下诸方面取得了较为明显的效果和不同程度的新进展。

1. 沉积盆地物源区的研究

沉积盆地中沉积物的组分多是各物源区母岩组分的混合。因子分析可以从沉积物样品的这种混合组分中识别出每个物源区的标准组分组合及其在沉积物样品组分中的比例，从而确定物源区的个数、母岩的类型以及各物源区对样品组分的影响，指出盆地周围物源区的分布。

2. 沉积物的粒度分析

沉积物的粒度参数包含着大量有关沉积搬运方式、腐蚀历史和水动力学等方面的信息，但它们一般又是多种水动力环境下的产物，因此应用因子分析就有可能找出一些典型的粒度组合，从而据此识别与沉积物相应的水动力环境。

3. 沉积相研究

因子分析可以把沉积岩的矿物成分、生物组合、岩石粒度、颜色、厚度、结构构造特征等在沉积相的意义下进行组合，从而启发我们识别各种沉积相及研究它们的空间分布。

4. 盆地生油条件和油气运移研究

根据生油岩样品抽提物和油样中与甾烷有关的地球化学分析数据，应用因子分析方法研究生油岩的母质类型、成熟度和烃的运移效应，对盆地的生油条件及含油气远景进行评价和油源对比工作。

5. 地层分析

对于巨厚的单一岩性的岩层来说，在它的剖面上往往很难看出沉积环境的明显变化。利用因子分析有可能认别出随着时间的推移在剖面上发生的气候、水深、物质来源和水动力条件等沉积环境因素的细微变化，从而使人们深化对这种地层剖面的认识。

6. 古环境与古生态的研究

古环境和古生态研究是古生物学研究中的具有高度成因性研究的课题。但对这二者却都不能用直接观察的方法来研究，而只能根据保存在岩石中的古生物化石组合和沉积物的地质、

物理和化学特征进行推论。因子分析可按照生态意义和环境意义对各种生物化石和沉积物特征进行组合。这种意义下的组合将会极大地启发古生物学家去思考诸如古气候、古温度、压力、PH 值、Eh 值、古盐度和古水深等古环境特征和古生物群对这种古环境相应的古生态反映。

7. *岩浆岩岩石化学成分的研究*

在岩浆岩岩石化学成分的研究中，应用因子分析可以帮助解决两个方面的问题：一是识别岩浆岩的形成过程，诸如岩浆的异源叠加、同源多期侵入、分异作用、同化作用、交代蚀变作用、矿化活动等；二是对岩浆岩分类。当然，这种分类结果必然带有因子分析所赋予的成因解释。

8. *变质岩的母岩性质*

因子分析既然可以根据沉积物的组分追溯沉积盆地的物源，那么同样可以用它来恢复形成变质岩的母岩。二者的不同之处在于物源区的研究中是识别在同一时间点不同空间过程的叠加，而变质岩母岩的恢复则是识别同一空间点上不同时间过程的叠加。

通过因子分析，可以识别出代表母岩特有的标准组分或条件的组合，从而达到恢复母岩的目的。

9. *矿床成因研究*

矿床学的研究与岩石学、沉积学、地层学和地球化学等各方面问题的研究是分不开的，因此，因子分析方法所能研究和解决的各种地质问题，在矿床学研究中都有可能碰到。从矿床学本身来讲，因子分析可以帮助解决两方面的问题：其一是识别矿化活动的阶段和类型；其二是分析成矿控制因素。矿床成因的恢复就依赖于这两个问题的解决。

10. *矿物的类质同象研究*

研究象角闪石和辉石这类矿物的类质同象，除测试手段外，因子分析也是一种简便易行的方法。因子分析可以提供两个元素或两个元素之间的类质同象替代模型，从而为研究元素的替代规律和晶体结构提供依据。

11. *地球化学研究*

在多数情况下，从岩石或矿物中获得的元素含量，实际上是在多个地质过程叠加条件下元素行为历史的总和。因子分析的任务不仅要根据岩石或矿物中各元素之间的相关关系来识别一个地质过程以及在该过程中元素的迁移富集规律，而且在多过程叠加的情况下，要区开分这些过程，并按不同的地质过程将元素的历史行为进行分解，因此，可以用因子分析将一个样品的观测值分解为背景和异常值两部分。背景值是一地质过程的产物，它同样品的形成过程有关，而异常值为另一局部过程的产物。

12. *水化学研究*

由于水是容易运移的流体，故地层水往往是不同水源的混合物，利用因子分析可以判断形成混合水的水型及分布，这对于寻找盐类矿床和油气田具有重要的意义。

§2　R 型因子分析

一、R 型因子分析的一般数学模型

（一）主因子

在一般情况下，地质变量是时间上和空间上不同地质过程叠加的结果。因此，它们之间往往具有比较复杂的关系，既有相关的一面，又有独立的一面。变量间的复杂关系就决定了地质研究时不宜直接研究单一的地质变量，最好是研究由它们的组合构成的少数几个综合变量，这

样的综合变量又叫主因子。主因子不仅具备相关性极小或不相关的特点，而且又能把原始地质变量所包含的不十分明显的差异尽可能多的反映出来。为了进一步说明主因子的概念，先看一个简单的例子。

设有 24 个样品，每个样品有 x_1 和 x_2 两个变量，它们的观测值见表 7-1，散点图如图 7-1 所示。

表 7-1 标准化变量观测值

样品号	变量 x_1	变量 x_2	样品号	变量 x_1	变量 x_2
1	−1.7984	−1.8020	13	0.1199	0.3713
2	−1.5586	−0.0634	14	0.3597	−0.0634
3	−1.0791	−1.1500	15	0.3597	0.1539
4	−1.0791	−0.4980	16	0.5995	−0.9327
5	−1.0791	−0.0634	17	0.5995	0.8059
6	−0.8393	−1.8020	18	0.5995	1.0233
7	−0.8393	0.5885	19	0.5995	1.4579
8	−0.5995	−0.2807	20	0.8393	−0.7154
9	−0.3597	−1.1500	21	1.0791	0.5885
10	−0.3597	−0.4980	22	1.5586	0.5886
11	−0.3597	0.8059	23	1.5586	1.4579
12	−0.1199	−0.7154	24	1.7984	1.8926

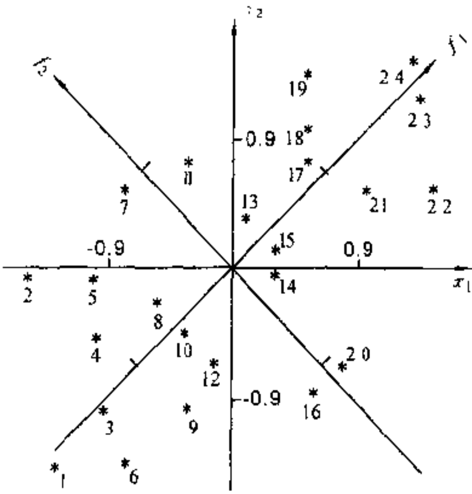


图 7-1 主因子示意图

如果把 x_1 轴和 x_2 轴同时按逆时针方向旋转一个角度 φ，得新的坐标轴 f_1 和 f_2。记

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

其中 $a_{11}=\cos\varphi, a_{21}=\sin\varphi, a_{12}=-\sin\varphi, a_{22}=\cos\varphi$。根据坐标旋转变换，变量 x_1、x_2 与变量 f_1、f_2 之间应有如下关系：

$$\begin{cases} f_1 = a_{11}x_1 + a_{21}x_2 \\ f_2 = a_{12}x_1 + a_{22}x_2 \end{cases} \quad 或 \quad \begin{cases} x_1 = a_{11}f_1 + a_{12}f_2 \\ x_2 = a_{21}f_1 + a_{22}f_2 \end{cases} \tag{7-3}$$

式(7-3)的矩阵形式为：

$$F = A'X \quad 或 \quad X = AF \tag{7-4}$$

设 $\varphi=44°$，根据式(7-3)可求得新变量 f_1 和 f_2 的值。见表 7-2。

表 7-2 新变量 f_1、f_2 数据

样品号	变量 f_1	变量 f_2	样品号	变量 f_1	变量 f_2
1	−2.5454	−0.0470	13	0.3442	0.1838
2	−1.1632	1.0371	14	0.2147	−0.2955
3	−1.5751	−0.0776	15	0.3657	0.1392
4	1.1222	0.3914	16	−0.2167	−1.0874
5	−0.8203	0.7040	17	0.9911	0.1633

续表 7-2

样品号	变量		样品号	变量	
	f_1	f_2		f_1	f_2
6	−1.8555	−0.7132	18	1.1421	0.3197
7	−0.1949	1.0064	19	1.4440	0.6323
8	−0.6262	0.2145	20	0.1068	−1.0976
9	−1.0576	−0.5774	21	1.1851	−0.3262
10	−0.6047	−0.1084	22	1.5300	−0.6593
11	0.3011	0.8296	23	2.1339	−0.0340
12	−0.5832	−0.4313	24	2.6084	0.1121

由表 7-2 数据可知，新变量 f_1 和 f_2 的平均值均等于 0，但两者的方差却不同。f_1 的方差为 1.6455，而 f_2 的方差等于 0.3535，f_1 的方差占总方差的 82%以上，这表明新变量 f_1 能反映原始数据的绝大部分信息。事实上，f_1 的最大值为 2.6048，最小值为 −2.5454，极差等于 5.1538，而 f_2 的最大值为 1.0371，最小值为 −1.0976，极差是 2.1347。这就表明 f_1 的波动范围是 f_2 波动范围的 2 倍还要多。在这种情况下，忽略 f_2 的影响也无损大局。在某些情况下，f_2 的变化范围也许会小于允许的观测误差，这时它就完全失去了存在的必要，因此就可用 f_1 代替 x_1 和 x_2 了。这样式(7-3)则可改写成如下形式：

$$f_1 = a_{11}x_1 + a_{21}x_2 \text{ 或 } x_i = a_{i1}f_1 + a_i e_i \quad (i = 1,2) \tag{7-5}$$

其中 e_i 服从均值为 0，方差为 σ_i^2 的正态分布。

如果每个样品有 m 个变量 $x_1, x_2, \cdots, x_m$，那么 m 个变量可以综合出 m 个新变量：

$$f_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{mj}x_m \quad (j = 1,2,\cdots,m) \tag{7-6}$$

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{im}f_m \quad (i = 1,2,\cdots,m) \tag{7-7}$$

由式(7-3)知，系数的平方和应满足

$$\sum_{k=1}^{m} a_{ik}^2 = 1 \quad (i = 1,2,\cdots,m)$$

在这种条件下，由原始变量线性组合而成的新变量 f_j 叫做综合变量或主因子。因子分析的任务之一就是找出 $p(p<m)$ 个主因子，把原始变量表示为 p 个主因子的线性组合，以此化简变量的研究系统。

（二）R 型因子分析模型

在式(7-6)中的系数为已知的条件下，可求得 n 个样品 m 个变量 f_1、f_2、$\cdots f_m$ 的值为

$$F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{bmatrix}$$

因此，n 个样品 m 个变量 $x_1, x_2, \cdots x_m$ 观测值写成矩阵形式为

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{bmatrix}$$

或者

$$X = AF \tag{7-8}$$

在进行综合研究时，如果用前 $p(p<m)$ 个主因子就能解释原始数据 80～90%以上的信息，那么式(7-8)可改写为

$$X = AF = [A_1\ A_2]\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = A_1F_1 + A_2F_2 \tag{7-9}$$

这里

$$A_1 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix} \qquad A_2 = \begin{bmatrix} a_{1p-1} & a_{1p+2} & \cdots & a_{1m} \\ a_{2p+1} & a_{2p+2} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{mp+1} & a_{mp+2} & \cdots & a_{mm} \end{bmatrix}$$

$$F_1 = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{p1} & f_{p2} & \cdots & f_{pn} \end{bmatrix} \qquad F_2 = \begin{bmatrix} f_{p+11} & f_{p+12} & \cdots & f_{p+1n} \\ f_{p+21} & f_{p+22} & \cdots & f_{p+2n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{m1} & f_{m2} & \cdots & f_{mn} \end{bmatrix}$$

式(7-9)表明，原始数据所包含的信息可分解为两部分。其中 A_1F_1 是由 p 个主因子所能解释的部分，而 A_2F_2 可称为残余部分。残余部分可表示为 $aE=(a_1e_1\ a_2e_2 \cdots\ a_me_m)'$，其中 $e_i(i=1,2,\cdots,m)$ 相互独立且服从 $N(0,1)$ 正态分布，因此有

$$X = A_1F_1 + aE \tag{7-10}$$

对于变量 x_i，它的线性表达式如下：

$$x_i = a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{ip}f_p + a_ie_i \tag{7-11}$$
$$(i = 1,2,\cdots,m)$$

通常称式(7-10)或式(7-11)为 R 型因子分析模型。

因子模型中，在各变量中共同出现的因子 $f_1,f_2,\cdots,f_p$ 叫做公因子。它们是相互独立的理论变量，可将其理解为 P 维空间中互相垂直的 p 个坐标轴。$e_1,e_2,\cdots,e_m$ 叫做特殊因子，它们是每个单一变量所特有的因子，各特殊因子之间以及它们与所有公因子之间都是相互独立的。

a_{ij}叫做因子载荷，是第 i 个变量在第 j 个公因子轴上的负荷。如果把 x_i 视为 p 维空间中的一个向量，那么 a_{ij}则是 x_i 在坐标轴 f_j 上的投影。矩阵 A_1 称为因子载荷矩阵。

二、因子模型中各个量的统计意义

假设因子模型是标准化的，即变量、公因子和特殊因子的平均值为 0，方差为 1。下面给出某些量的统计意义。

1. 因子载荷的统计意义

对式(7-11)两边同时右乘 f_j，则有

$$x_if_j = a_{i1}f_1f_j + a_{i2}f_2f_j + \cdots + a_{ij}f_jf_j + \cdots + a_{ip}f_pf_j + a_ie_if_j \tag{7-12}$$

对式(7-12)两边取数学期望，即

$$\begin{aligned} E(x_if_j) &= a_{i1}E(f_1f_j) + a_{i2}E(f_2f_j) \\ &\quad + \cdots + a_{ij}E(f_jf_j) + \cdots + a_{ip}E(f_pf_j) + a_iE(e_if_j) \end{aligned} \tag{7-13}$$

由于模型是标准化的，故上式中的数学期望就是相关系数，因此有

$$r_{x_if_j} = a_{i1}r_{f_1f_j} + a_{i2}r_{f_2f_j} + \cdots + a_{ij}r_{f_jf_j} + \cdots + a_{ip}r_{f_pf_j} + a_ir_{e_if_j} \tag{7-14}$$

由公因子和特殊因子的独立性可得

$$\begin{cases} r_{f_i f_j} = 1 & (i = j) \\ r_{f_i f_j} = 0 & (i \neq j) \end{cases}$$

所以

$$r_{x_i f_j} = a_{ij} \tag{7-15}$$

式(7-15)表明,因子载荷 a_{ij}是第 i 个变量 x_i 与第 j 个公因子 f_j 的相关系数。

2. 公因子的方差

(1) 诸公因子方差

由式(7-11)可求得变量 x_i 的方差为

$$D(x_i) = a_{i1}^2 D(f_1) + a_{i2}^2 D(f_2) + \cdots + a_{ip}^2 D(f_p) + a_i^2 D(e_i)$$

根据因子模型标准化的假设条件,则有

$$D(x_i) = a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 + a_i^2 = 1, \quad 即$$

$$D(x_i) = \sum_{j=1}^{p} a_{ij}^2 + a_i^2 = h_i^2 + a_i^2 = 1 \tag{7-16}$$

式(7-16)表明,变量 x_i 的方差由两部分组成,它的第 1 部分 h_i^2 是因子载荷矩阵中第 i 行元素的平方和,是全部公因子对变量 x_i 的方差所做的贡献,将其简称为诸公因子方差;第 2 部分 a_i^2 是特殊因子对变量 x_i 的方差所做的贡献,是对变量 x_i 方差的补充值,将其简称为特殊因子方差。

(2) 公因子 f_j 的方差

因子载荷矩阵中列元素平方和

$$s_j = \sum_{i=1}^{m} a_{ij}^2 \qquad (j = 1, 2, \cdots, p)$$

的统计意义与诸公因子方差相反,它是同一个公因子 f_j 对诸变量所提供的方差的总和,将其简称为公因子 f_j 的方差,它是衡量公因子 f_j 相对重要性的一个指标。

三、因子载荷的几何意义

在 R 型因子模型的假设条件下,可把 p 个公因子和 m 个特殊因子视为$(p+m)$维空间中相互垂直的单位向量。这样,就由它们共同构成一个$(p+m)$维的直角坐标系,并把该坐标系称为因子空间。变量 x_i 就是因子空间中的一个向量,因子载荷 $a_{ik}(k=1,2,\cdots,p)$,$a_{jk}(k=1,2,\cdots,p)$,a_i、a_j 则是向量 x_i 和 x_j 在各因子轴上的投影。在这种情况下,变量 x_i 与 x_j 的相关系数就是它们的相似系数,即

$$\begin{aligned} r_{x_i x_j} &= \cos(x_i, x_j) = x_i \cdot x_j / (|x_i||x_j|) \\ &= a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{ip}a_{jp} + a_i a_j \\ &= \sum_{k=1}^{p} a_{ik}a_{jk} + a_i a_j \\ &= \begin{cases} \sum_{k=1}^{p} a_{ik}a_{jk} & (i \neq j) \\ h_i^2 + a_i^2 & (i = j) \end{cases} \end{aligned} \tag{7-17}$$

根据式(7-16),则有

$$r_{x_ix_j}=\begin{cases}\sum_{k=1}^{p}a_{ik}a_{jk} & (i\neq j)\\ 1 & (i=j)\end{cases} \tag{7-18}$$

§3　主因子的解

主因子的解是指因子模型中的因子载荷矩阵 A_1。在这一节中，讨论的问题是如何求出因子载荷矩阵。

一、约相关矩阵

根据式(7-17)，m 个标准化变量的相关矩阵可以写成如下形式：

$$R=\begin{pmatrix} h_1^2+a_1^2 & \sum_{k=1}^{p}a_{1k}a_{2k} & \cdots & \sum_{k=1}^{p}a_{1k}a_{mk}\\ \sum_{k=1}^{p}a_{2k}a_{1k} & h_2^2+a_2^2 & \cdots & \sum_{k=1}^{p}a_{2k}a_{mk}\\ \cdots & \cdots & \cdots & \cdots\\ \sum_{k=1}^{p}a_{mk}a_{1k} & \sum_{k=1}^{p}a_{mk}a_{2k} & \cdots & h_m^2+a_m^2\end{pmatrix}$$

$$=\begin{pmatrix} h_1^2 & \sum_{k=1}^{p}a_{1k}a_{2k} & \cdots & \sum_{k=1}^{p}a_{1k}a_{mk}\\ \sum_{k=1}^{p}a_{2k}a_{1k} & h_2^2 & \cdots & \sum_{k=1}^{p}a_{2k}a_{mk}\\ \cdots & \cdots & \cdots & \cdots\\ \sum_{k=1}^{p}a_{mk}a_{1k} & \sum_{k=1}^{p}a_{mk}a_{2k} & \cdots & h_m^2\end{pmatrix}+\begin{pmatrix} a_1^2 & 0 & \cdots & 0\\ 0 & a_2^2 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & a_m^2\end{pmatrix}$$

$$=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p}\\ a_{21} & a_{22} & \cdots & a_{2p}\\ \cdots & \cdots & \cdots & \cdots\\ a_{m1} & a_{m2} & \cdots & a_{mp}\end{pmatrix}\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1}\\ a_{12} & a_{22} & \cdots & a_{m2}\\ \cdots & \cdots & \cdots & \cdots\\ a_{1p} & a_{2p} & \cdots & a_{mp}\end{pmatrix}+\begin{pmatrix} a_1 & & & 0\\ & a_2 & & \\ & & \ddots & \\ 0 & & & a_m\end{pmatrix}\begin{pmatrix} a_1 & & & 0\\ & a_2 & & \\ & & \ddots & \\ 0 & & & a_m\end{pmatrix}$$

$=A_1A'_1+aa'$，记

$$R^*=R-aa'=A_1A'_1 \tag{7-19}$$

并称 R^* 为约相关矩阵。它与相关矩阵 R 的差别仅在于：R 的对角线元素均为 1，而 R^* 的对角线元素依次为诸公因子的方差。由式(7-19)可知，R 型因子分析就是在已知约相关矩阵 R^* 的条件下，求因子载荷矩阵 A_1，使得 $R^*=A_1A'_1$，这也就是求主因子解的基本出发点。

需要指出的是：如果已知约相关矩阵 R^*，并且 A_1 又是式(7-19)的解，那么对于任意一个 $p\times p$ 的正交矩阵 B，则有

$$(A_1B)(A_1B)'=(A_1B)(B'A'_1)=A_1BB'A'_1=A_1A'_1$$

这一结果表明，在 A_1 是式(7-19)解的情况下，A_1B 也必然是式(7-19)的解，即式(7-19)的解 A_1 并不是唯一的，它将随着附加条件的不同而不同。另外，在一般情况下，也不容易求得诸公因子对变量 x_i 的方差 h_i^2，因此也就不能确定变量的约相关矩阵 R^*。鉴于上述原因，下面介绍确定

A_1 的条件极值法。

二、求主因子解的条件极值法

这种方法的基本思想是先从变量相关矩阵 R 中选出一个公因子 f_1，并使它对诸变量所提供的方差为最大。选出第一个公因子 f_1 后，从 R 中扣除 f_1 的影响得到剩余相关矩阵 R_1，再从 R_1 中选出第 2 个公因子 f_2，并使 f_2 对诸变量提供的方差为最大，这样不断地选取 f_3、f_4…，直到把诸变量的方差分解完为止。

挑选第一个公因子 f_1，也就是要确定因子载荷 a_{i1}，使 f_1 对诸变量所提供的方差

$$s_1 = a_{11}^2 + a_{21}^2 + \cdots + a_{m1}^2 \tag{7-20}$$

在满足

$$r_{ij} = \sum_{k=1}^{p} a_{ik}a_{jk} \quad (i,j=1,2,\cdots,m) \tag{7-21}$$

的条件下达到最大。其中 $r_{ij}=r_{ji}$，r_{ii}是变量 x_i 的诸公因子方差 h_i^2。在式(7-21)条件下求函数 s_1 的极大值是一个求条件极值的问题，而求条件极值的常用方法是拉格朗日乘数法。令

$$T = s_1 - \sum_{i,j=1}^{m} \mu_{ij}r_{ij} = s_1 - \sum_{i,j=1}^{m}\sum_{k=1}^{p} \mu_{ij}a_{ik}a_{jk} \tag{7-22}$$

式(7-22)中 $\mu_{ij}=\mu_{ji}$为拉格朗日乘数，T 为 $a_{i1}(i=1,2,\cdots,p)$的函数。

求 T 对每一个变量 a_{i1}的偏导数，并令其等于 0，即

$$\frac{\partial T}{\partial a_{i1}} = a_{i1} - \sum_{j=1}^{m} \mu_{ij}a_{j1} = 0 \tag{7-23}$$

同样，也可以求 T 对其它每个变量 $a_{ik}(k\neq 1)$的偏导数，并令其等于 0，即

$$\frac{\partial T}{\partial a_{ik}} = -\sum_{j=1}^{m} \mu_{ij}a_{jk} = 0 \quad (k\neq 1) \tag{7-24}$$

把式(7-23)和式(7-24)合并为式(7-25)。

$$\frac{\partial T}{\partial a_{ik}} = \delta_{1k}a_{i1} - \sum_{j=1}^{m} \mu_{ij}a_{jk} = 0 \quad (k=1,2,\cdots,p) \tag{7-25}$$

式(7-25)中 $\delta_{1k}=\begin{cases}1 & (k=1)\\ 0 & (k\neq 1\end{cases}$。用 a_{i1}乘以式(7-25)的两边并对 i 求和，得到

$$\delta_{1k}\sum_{i=1}^{m} a_{i1}^2 - \sum_{i=1}^{m}\sum_{j=1}^{m} \mu_{ij}a_{i1}a_{jk} = 0 \tag{7-26}$$

由式(7-23)得

$$\sum_{i=1}^{m} \mu_{ji}a_{i1} = a_{j1}$$

因此，式(7-26)变为

$$\delta_{1k}s_1 - \sum_{j=1}^{m} a_{j1}a_{jk} = 0 \tag{7-27}$$

再用 a_{ik}乘以式(7-27)的两边并对 k 求和，则有

$$a_{i1}s_1 - \sum_{j=1}^{m} a_{j1}\left(\sum_{k=1}^{p} a_{ik}a_{jk}\right) = 0 \quad (i=1,2,\cdots,m)$$

应用式(7-21)附加条件，有

$$\sum_{j=1}^{m} r_{ij}a_{j1} - s_1a_{i1} = 0 \quad (i=1,2,\cdots,m)$$

写成矩阵形式为：

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} - s_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{7-28}$$

记 $a_1=(a_{11}a_{21}\cdots a_{m1})'$，$0=(00\cdots0)'$，$E$ 为 $m\times m$ 阶单位矩阵，式(7-28)化简为

$$(R - S_1E)a_1 = 0 \tag{7-29}$$

其中 R 为相关矩阵，$a_{i1}(i=1,2,\cdots,m)$为因子载荷。主因子要求 $a_{11},a_{21}\cdots a_{m1}$不能同时为 0，因此，要使齐次线性方程组(7-29)有非零解，那么它的系数行列式必须等于 0，即

$$|R - S_1E| = 0 \tag{7-30}$$

式(7-30)为相关矩阵 R 的特征方程。由于要求 S_1 为最大，所以 S_1 应等于 R 的最大特征值 λ_1，即

$$S_1 = \sum_{i=1}^{m} a_{i1}^2 = \lambda_1 \tag{7-31}$$

把式(7-31)代入式(7-29)并解这一线性方程组，可求得一组解 $\alpha_{11},\alpha_{21},\cdots,\alpha_{m1}$。这一组解称为对应于特征值 λ_1 的特征向量。特征向量 α_{i1}与因子载荷 a_{i1}有如下关系：

设 R^* 为约相关矩阵；Λ 为对角矩阵，其对角元素为 λ_i；P 为正交方阵；$M_k=P_1P_2\cdots P_k$；M'_k 为 M_k 的转置；可以证明

$$M'_kR^*M_k = \Lambda \tag{7-32}$$

即对约相关矩阵 R^*进行足够多次(k 次)正交变换后，R^* 则成为以 λ_i 为对角元的对角矩阵 Λ；而 M_k 则为 R^* 的特征向量所构成的矩阵。对式(7-32)均左乘 M_k，右乘 M'_k，得

$$R^* = M_k\Lambda M'_k$$

据式(7-19)，则有

$$A_1A'_1 = M_k\Lambda^{1/2}\Lambda^{1/2}M'_k$$

其中 $\Lambda^{1/2}$为以$\sqrt{\lambda_i}$ 为对角元的对角矩阵。从而

$$A_1=M_k\Lambda^{1/2}$$

因此公因子 f_1 的载荷应为

$$a_{i1} = \alpha_{i1}\sqrt{\lambda_1} \tag{7-33}$$

α_{i1}是与 R^*的特征值 λ_1 对应的特征向量。

选出第 1 个公因子 f_1 之后，如果诸变量的公因子方差没有被分解完，这时就要继续选择第 2 个公因子 f_2，它与 f_1 互不相关，而且它对诸变量提供的方差总和

$$S_2 = a_{12}^2 + a_{22}^2 + \cdots + a_{m2}^2$$

在满足条件

$$R^{(1)} = R - a_1a'_1 \tag{7-34}$$

或者

$$r_{ij}^{(1)} = r_{ij} - a_{i1}a_{j1} = \sum_{k=2}^{p} a_{ik}a_{jk} \quad (i,j=1,2,\cdots,m)$$

下为最大。$R^{(1)}$是从 R 中扣除公因子 f_1 的影响之后的剩余相关矩阵，$r_{ij}^{(1)}$为 $R^{(1)}$中第 i 行第 j 列

的元素。选择方法同前，即先用拉格朗日乘数法求条件极值，然后再利用特征方程确定公因子f_2的因子载荷。

由上述方法确定每一个公因子时，都必须重新计算剩余相关矩阵，当需要确定的公因子数较多时，计算过程较复杂，并且计算量也很大。事实上，$R^{(1)}$的最大特征值却是原始相关矩阵R的次大特征值。若能证明这一点，则可加快整个求解过程。关于这个问题，只要证明R的P个特征向量同时也是$R^{(1)}$的特征向量即可。

设$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$是相关矩阵$R$的$P$个特征值，$\alpha_1, \alpha_2, \cdots, \alpha_p$是与$\lambda_i (i=1,2,\cdots,p)$对应的$P$个规格化特征向量。对于任何的$k=1,2,\cdots,p$，根据式(7-29)有

$$R\alpha_k = \lambda_k \alpha_k$$

以α_k右乘式(7-34)，则有

$$R^{(1)}\alpha_k = (R - a_1 a'_1)\alpha_k = R\alpha_k - a_1 a'_1 \alpha_k = \lambda_k \alpha_k - a_1 a'_1 \alpha_k \tag{7-35}$$

在式(7-35)中，当$k=1$时，根据式(7-31)得

$$R^{(1)}\alpha_1 = 0 \tag{7-36}$$

式(7-35)表明，对应于相关矩阵R中的最大特征值λ_1的特征向量也是$R^{(1)}$的一个特征向量，只是在$R^{(1)}$中与之对应的特征值为0。当$k \neq 1$时，由于$a_1 a'_1 = 0$，故式(7-35)化简为

$$R^{(1)}\alpha_k = \lambda_k \alpha_k \quad (k \neq 1) \tag{7-37}$$

式(7-37)表明，除相关矩阵R的最大特征值λ_1外，剩余相关矩阵$R^{(1)}$的特征值和R的特征值相同，并且两者对应的特征向量也相同。

由上述可知：R的次大特征值是$R^{(1)}$的最大特征值。这就是说，我们可以不必计算剩余相关矩阵，而直接对原始相关矩阵R进行特征值分析，取相关矩阵R的特征值λ_2、λ_3、…、λ_p及对应的特征向量α_2、α_3、…、α_p依次作为剩余相关矩阵$R^{(1)}$、$R^{(2)}$、…、$R^{(p-1)}$的最大特征值和特征向量，即可得因子载荷矩阵

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix} = \begin{bmatrix} \alpha_{11}\sqrt{\lambda_1} & \alpha_{12}\sqrt{\lambda_2} & \cdots & \alpha_{1p}\sqrt{\lambda_p} \\ \alpha_{21}\sqrt{\lambda_1} & \alpha_{22}\sqrt{\lambda_2} & \cdots & \alpha_{2p}\sqrt{\lambda_p} \\ \vdots & \vdots & & \vdots \\ \alpha_{m1}\sqrt{\lambda_1} & \alpha_{m2}\sqrt{\lambda_2} & \cdots & \alpha_{mp}\sqrt{\lambda_p} \end{bmatrix}$$

综上所述，在R型因子分析中，求主因子的解就是求约相关矩阵的特征值和与之对应的特征向量。

§4　方差最大正交旋转

通过因子分析可以找出主要的公因子，但是，更重要的是要明确每个因子的实际地质意义。为此，还需要对因子载荷施行旋转使其结构化，使每个因子载荷的平方向1或0两极分化，其中的第j个公因子的代表性变量在f_j因子轴上的载荷系数等于或趋近于1，而在其他因子轴上的载荷系数等于或趋近于0。这样便容易对每个公因子进行地质解释。此种作法，从数学上来说就是对矩阵A_1进行正交变换，亦即将各因子轴在它们所确定的空间中作一正交旋转。

目前，对因子载荷的旋转方法有许多种，但其中最常用方法是方差最大正交旋转，这种旋转方法是使因子载荷矩阵中的各因子载荷值的方差达到最大。对于R型因子载荷矩阵$A_1=$

$[a_{ij}]_{m\times p}$，对因子 f_j 的简化，可由因子载荷值的平方方差来表示，即

$$V_j = \frac{1}{m}\sum_{i=1}^{m}(b_{ij}^2)^2 - \left(\frac{1}{m}\sum_{i=1}^{m}b_{ij}^2\right)^2 = \left[m\sum_{i=1}^{m}(b_{ij}^2)^2 - \left(\sum_{i=1}^{m}b_{ij}^2\right)^2\right]/m^2$$

式中的 b_{ij}是经过正交旋转后所得因子载荷矩阵 B 的元素。使用载荷值的平方是为了避免出现负值。

如果使 V_j 为最大，亦即使第 j 个因子得到最大的简化，此时，它在因子空间中 f_j 的载荷系数趋于 1，而在其它因子轴上的载荷系数趋于 0，那么，对于整个因子载荷矩阵 $A_1=[a_{ij}]_{m\times p}$ 的简化则可由所有因子载荷的平方方差之和作为衡量标准，即使

$$V = \sum_{j=1}^{p}V_j = \sum_{j=1}^{p}\left[m\sum_{i=1}^{m}(b_{ij}^2)^2 - \left(\sum_{i=1}^{m}b_{ij}^2\right)^2\right]/m^2 \tag{7-38}$$

达到最大。考虑到各个变量 $x_i(i=1,2,\cdots,m)$的诸公因子方差之间的差异，需要用(b_{ij}/h_{ij}^2)来代替式(7-38)中的 b_{ij}^2，这实际上是需要求得经过旋转后的 b_{ij}，使其

$$V = \sum_{j=1}^{p}\left[m\sum_{i=1}^{m}(b_{ij}^2/h_{ij}^2)^2 - \left(\sum_{i=1}^{m}(b_{ij}^2/h_{ij}^2)\right)^2\right]/m^2 \tag{7-39}$$

达到最大。

对因子载荷矩阵 $A_1=[a_{ij}]_{m\times p}$进行正交旋转，相当于对所有因子平面 $f_gf_q(g=1,2,\cdots,m-1;q=g+1,g+2,\cdots,m)$正交旋转一个角度 φ_{gq}，每次的旋转角 φ 必须满足使式(7-39)中的 V 达到最大值。为此，可选择如下的正交变换

$$T_{gq} = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & \cos\varphi & & -\sin\varphi & & \\ & & & \ddots & & & \\ & & \sin\varphi & & \cos\varphi & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}_{m\times m} \begin{matrix} \\ \\ g \\ \\ q \\ \\ \\ \end{matrix}$$

$$\qquad\qquad\qquad g \qquad\qquad\qquad q$$

T_{gq}中凡没有标明的元素均为 0，A_1 经过变换后，相当于将因子平面 f_gf_q 旋转一个角度 φ，得到矩阵

$$B = A_1T_{gq} = [b_{ij}]_{m\times p}$$

B 中的元素分别为

$$\begin{aligned} b_{ig} &= a_{ig}\cos\varphi + a_{iq}sin\ \varphi \\ b_{iq} &= -a_{ig}\cos\varphi + a_{iq}\cos\varphi \\ b_{ik} &= a_{ik} \qquad (k\neq g,q;i=1,2,\cdots,m) \end{aligned} \tag{7-10}$$

如果有 p 个主因子，则必须对 A_1 中所有 p 列全部配对旋转，总共旋转

$$C_p^2 = \frac{p(p-1)}{2}$$

次，全部旋转完毕算一个循环，此时，得到载荷矩阵

$$B_1 = A_1T_{12}\cdots T_{1p}\cdots T_{(p-1)p} = A_1\prod_{g=1}^{p-1}\prod_{q=g+1}^{p}T_{gq} = A_1C_1$$

上式中，$C_1=\prod_{g=1}^{p-1}\prod_{q=g+1}^{p}T_{gq}$，$B_1$ 为对 A_1 进行正交变换 C 而得。

经过第一个循环后，可按式(7-39)计算 V_1。在第一个循环基础上，从 B_1 出发进行第二个旋转循环，旋转完成后得到 B_2，即

$$B_2=B_1\prod_{g=1}^{p-1}\prod_{q=g+1}^{p}T_{gq}=B_1C_2=A_1C_1C_2$$

由 B_2 可计算出 V_2。

如此不断地重复这个步骤，就可以得到 V 的一个非降有界序列

$$V_1\leqslant V_2\leqslant\cdots$$

众所周知，由于因子载荷的绝对值不大于 1，所以这个序列是有上界的，必然收敛于某一极限 $V_{\max}$。$V_{\max}$为 V 的极大值，因此，只要循环次数 k 充分大，就必然有

$$|V_k-V|<\varepsilon$$

ε 为所要求的计算精度。如果循环次数 k 与$(k+1)$都充分大时，也必然有

$$|V_k-V_{k+1}|<\varepsilon$$

最后得

$$B_k=A_1\prod_{i=1}^{k}C_i=A_1C$$

B_k 就是旋转后的因子载荷矩阵。

前已述及，在任何一次变换 T_{gq} 中，必须使方差达到极大，为此，应按如下步骤确定旋转角度：

(1) 将式(7-40)代入式(7-39)；

(2) 将式(7-39)对 φ 求一阶导数并令其为 0，可解得

$$\operatorname{tg}4\varphi=\frac{D-2A_1B/m}{C-(A_1^2-V^2)/m}=\frac{E}{F} \tag{7-41}$$

上式中的 m 为变量个数。令

$$T_j=(a_{jg}/h_j)^2-(a_{jq}/h_j)^2$$
$$H_j=2(a_{jg}/h_j)(a_{jq}/h_j)$$

则
$$A_1=\sum_{j=1}^{m}T_j \qquad B=\sum_{j=1}^{m}H_j$$
$$C=\sum_{j=1}^{m}(T_j^2-H_j^2), \qquad D=2\sum_{j=1}^{m}T_jH_j$$

(3) 将式(7-39)展开，并将包含 φ 的项合并简化，最后就只剩下包含 $\sin 4\varphi$ 和 $\sin^2 2\varphi$ 的项，使式(7-39)成为以$\frac{\pi}{2}$为周期的函数，因而，式(7-41)中的 4φ 只要在$\frac{\pi}{2}$的范围内考虑就行。通常在$-\frac{\pi}{4}\sim\frac{\pi}{4}$之间考虑，同时由式(7-39)对 φ 的二阶导数应小于 0，可得

$$\frac{1}{E}\sin 4\varphi>0$$

所以，φ 的符号可根据 E 的符号确定，它应与 E 同号，所以，可按分子 E 及分母 F 的正负号来确定 4φ 应在哪一象限中。

以上以 R 型因子分析为例，对方差最大正交旋转问题进行了讨论。对于 Q 型因子分析，载荷矩阵的最大正交旋转情况相仿，Q 型因子分析 $A_1=[a_{ij}]_{n\times p}$，所以，只要将以上公式中的 m

换成 n 即可。

§5 因子得分

如前所述，主因子 f_j 是由原始变量 $x_1,x_2,\cdots x_m$ 线性组合而成的综合变量，即

$$f_j = c_{1j}x_1 + c_{2j}x_2 + \cdots + c_{mj}x_m \tag{7-42}$$
$$(j = 1,2,\cdots,p)$$

而这里的因子得分就是把第 i 个样品 m 个变量的观测值 $(x_{1i},x_{2i},\cdots x_{mi})$ 代入式(7-42)而计算出的函数值 $f_{ji}(j=1,2,\cdots,p;i=1,2,\cdots,n)$。

把式(7-42)写成矩阵形式，有

$$F = CX \tag{7-43}$$

其中 $F=(f_1\ f_2\cdots\ f_p)'$，$X=(x_1\ x_2\cdots\ x_m)'$，而

$$C = \begin{bmatrix} c_{11} & c_{21} & \cdots & c_{m1} \\ c_{12} & c_{22} & \cdots & c_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ c_{1p} & c_{2p} & \cdots & c_{mp} \end{bmatrix}$$

欲求因子得分，必须要先确定式(7-43)中的系数矩阵 C。

在因子分析模型式(7-10)中，若因子载荷矩阵 A_1 为满秩的 m 阶方阵，这时有 $ae=0$，$A_1=A$，于是，由式(7-10)可直接得 $F=A^{-1}X$，即 $C=A^{-1}$。如果 A_1 不是满秩的方阵，而是 $(m\times p)$ 阶的长方阵 $(m>p)$，且假定 $ae\approx 0$，那么有

$$X \approx A_1F \tag{7-44}$$

先对式(7-44)左乘 A'_1 后，再左乘 $(A'_1A_1)^{-1}$，则有

$$F = (A'_1A_1)^{-1}A'_1X \tag{7-45}$$

由式(7-43)和式(7-45)，得

$$C = (A'_1A_1)^{-1}A'_1 \tag{7-46}$$

一般说来，ae 将随着所选取因子数的减少而变大，当它变得不可忽略时，用式(7-46)求出的系数就不能计算出正确的因子得分。在这种情况下，只能在最小二乘法的意义下对因子得分进行估计。为此，必须建立因子 $f_j(j=1,2,\cdots,p)$ 对变量 $x_i(i=1,2,\cdots,m)$ 的回归方程。在因子分析模型已标准化的条件下，设因子 f_j 对 $x_i(i=1,2,\cdots,m)$ 的回归方程为

$$\hat{f}_j = b_{1j}x_1 - b_{2j}x_2 + \cdots + b_{mj}x_m \tag{7-47}$$
$$(j = 1,2,\cdots,p)$$

那么 p 个回归方程的矩阵形式为

$$\hat{F} = BX \tag{7-48}$$

式(7-48)中，$p\times n$ 阶矩阵 $\hat{F}$ 是矩阵 F 的最小二乘解，$p\times m$ 阶矩阵 B 是回归方程的系数矩阵，X 为 $m\times n$ 阶的原始数据矩阵。

分别用 $\frac{1}{n-1}X'$ 右乘式(7-48)两边，得

$$\frac{1}{n-1}\hat{F}X' = BXX'/(n-1)$$

因为$\frac{1}{n-1}XX'$为变量间的相关矩阵 R，而$\frac{1}{n-1}\hat{F}X'$是公因子与变量的相关矩阵 A'_1，所以上式可写成

$$A'_1 = BR$$

从上式中可解出 B

$$B = A'_1R^{-1}$$

把 B 代入式(7-48)，最后得到所要求的解

$$\hat{F} = A'_1R^{-1}X \tag{7-49}$$

式(7-49)中

$$\hat{F} = (\hat{f}_1\ \hat{f}_2 \cdots \hat{f}_p)',\quad X = (x_1\ x_2 \cdots x_m)'$$

$$R = \begin{bmatrix} r_1 & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \qquad A'_1 = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1p} & a_{2p} & \cdots & a_{mp} \end{bmatrix}$$

§6 Q型因子分析

R 型因子分析方法是在样品的基础上研究变量之间的相互关系，而 Q 型因子分析方法则是在变量的基础上研究样品之间的相互关系。变量之间的相互关系表现在原始数据矩阵 X 的行之间，而样品之间的相互关系则表现在同一矩阵的列之间，因此，在进行 Q 型因子分析时，需要把在 R 型因子分析中的变量和样品的下标调换过来。

一、相似矩阵

衡量样品之间相似性的度量之一是相似系数，即 m 维空间中两个样品点向量 $x_i=(x_{1i}\ x_{2i}\cdots x_{mi})'$ 与 $x_j=(x_{1j}\ x_{2j}\cdots x_{mj})'$ 之间夹角 θ_{ij} 的余弦，记为 q_{ij}：

$$q_{ij} = \cos\theta_{ij} = \frac{\sum_{k=1}^{m} x_{ki}\cdot x_{kj}}{\sqrt{\sum_{k=1}^{m} x_{ki}^2 \sum_{k=1}^{m} x_{kj}^2}} \qquad (i,j=1,2,\cdots,n) \tag{7-50}$$

这里：x_{ki} 是第 i 个样品第 k 个变量的观测值；m 为变量数；n 为样品数。

n 个样品之间的相似系数构成一个$(n\times n)$的相似矩阵

$$Q = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{bmatrix}$$

在 Q 型因子分析中，常对样品进行模标准化，即用每一样品向量的长度去除 X 的相应列。由于这样做并不改变样品中各个变量的比例关系，因此，样品之间的相似系数仍保持不变，也就是说，不会影响分析的结果。对变换后的数据来说，由于每一列的平方和为 1，因此，在样品空间中样品向量都具有单位长度，向量的端点就都在单位超球面上。

当变量的量纲不同，导致样品的观测值的数量级差异明显时，往往先对原始数据矩阵的行

作极差标准化，然后再对极差标准化后的数据矩阵的列进行模标准化。

二、Q 型因子分析模型

Q 型因子分析是从样品的相似矩阵出发，研究样品之间的相互关系以及探索样品产生相似性的原因。除此之外，一切都与 R 型因子分析相类似，因此，它也有与 R 型因子分析相类似的数学模型。

如果由 p 个因子（综合变量）$f_k(k=1,2,\cdots,p)$ 可以把给定的 n 个样品 $x_1,x_2\cdots,x_n$ 线性表出，即

$$x_j = a_{j1}f_1 + a_{j12}f_2 + \cdots + a_{jp}f_p + a_je_j \quad (j=1,2,\cdots,n) \tag{7-51}$$

其中 $a_{ji}(i=1,2,\cdots,p;p\leqslant m)$ 是第 j 个样品 x_j 在因子 f_i 上的载荷。称式(7-51)为 Q 型因子分析模型。

三、主因子解

Q 型因子模型也有与 R 型因子模型相似的主因子解：

$$a_{ij} = u_{ij}\sqrt{\lambda_j} \qquad (i=1,2,\cdots,n;j=1,2,\cdots,p) \tag{7-52}$$

u_{ij} 是与相似矩阵的特征值 λ_j 对应的特征向量。

四、因子得分

在 Q 型因子分析模型中，当 ae 不可忽略时，在最小二乘法的意义下也可得到类似于式(7-49)的因子得分计算公式。但 Q 型因子分析中诸公因子的方差收敛很快，故常用式(7-45)计算因子得分。

§7 对应分析

对应分析是在 R 型因子分析和 Q 型因子分析基础上发展起来的一种多元统计分析方法，它把两种因子分析结合起来，对变量和样品统一进行分析研究，因而更有利于地质解释。

如前所述，两种因子分析都可以用少数几个公因子去提取研究对象的绝大部分信息，因而，不仅化简了原有的观测系统，抓住了控制原有观测数据的主要矛盾，而且通过研究公因子的特征，比较容易揭示研究对象在成因上或空间上的联系，也就便于直接进行地质解释和逻辑推断。但是，R 型因子分析与 Q 型因子分析把变量与样品孤立起来分析，割断了它们的联系，这将会漏掉许多有用的地质信息。事实上，对于同一个地质问题，往往需要同时研究地质成因和不同类型样品的地质特征，前者要通过对样品的研究，而后者则是通过对变量的分析，才能得到合理的地质解释。这说明两种因子分析是同一问题的不可分割的两个部分。另外，样品的数目一般远远大于变量的数目，在进行 Q 型因子分析时，样品的相似矩阵占用大量的内存，这对于一般的微型计算机来说是难以胜任的。还有一个问题就是不能对变量和样品用同一种标准化方法进行处理，这就给寻找 R 型与 Q 型因子分析之间的联系带来了困难。

鉴于上述原因，在 R 型因子分析和 Q 型因子分析的基础上产生了对应分析。它的主要优点是可由 R 型因子分析的结果，很容易地导出 Q 型因子分析结果，从而克服了 Q 型因子分析受计算机内存容量的限制并提高了计算速度，更重要的是把变量和样品反映在同一个因子空间中，便于对变量与样品统一进行地质解释和推断。

一、原始数据的变换

设有 n 个样品，每个样品有 m 个变量，它们的原始观测值记为

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

其中 $x_{ij} \geqslant 0 (i=1,2,\cdots,m; j-1,2,\cdots,n)$，并且在每一行和每一列上至少有一个不为 0 的数。

记数据矩阵 X 第 i 行元素之和为

$$x_{i\cdot} = \sum_{k=1}^{n} x_{ik} \quad (i = 1,2,\cdots,m)$$

第 j 列元素之和为

$$x_{\cdot j} = \sum_{k=1}^{m} x_{kj} \quad (j = 1,2,\cdots,n)$$

而 X 中 $(m \times n)$ 个元素之和为

$$T = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$$

并用 T 去除 X 的每个元素，得一个新的数据矩阵，记为 P

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{pmatrix}$$

再用 P 矩阵第 j 列的和

$$P_{\cdot j} = \sum_{i=1}^{m} p_{ij} \qquad (j = 1,2,\cdots,n)$$

去除以 P 矩阵中第 j 列上的每个元素，得

$$\left(\frac{p_{1j}}{p_{\cdot j}}\ \frac{p_{2j}}{p_{\cdot j}} \cdots \frac{p_{mj}}{p_{\cdot j}}\right)' \quad (j = 1,2,\cdots,n) \tag{7-53}$$

在 m 维空间中，用以式(7-53)为坐标的 n 个点表示 n 个样品，称它们为样品点，每个样品点的坐标是各个变量在该样品中的相对比例，因此，研究 n 个样品的相似性就化为研究 m 维空间中 n 个样品点的相对位置。两个样品点相距越近，那么它们的性质就越相似。

在研究样品点的相对位置时，如果第 i 个变量数量级较大，那么它对两样品点间的距离就有较大 的影响。为了消除由于量纲不同导致数量级对距离的影响，可以采用加权距离 $D(l,k)$ 作为两个样品点 l 和 k 之间接近程度的度量

$$D(l,k) = \sqrt{\sum_{i=1}^{m} \frac{1}{p_{i\cdot}}\left(\frac{p_{il}}{p_{\cdot l}} - \frac{p_{ik}}{p_{\cdot k}}\right)^2} = \sqrt{\sum_{i=1}^{m}\left(\frac{p_{il}}{p_{\cdot l}\sqrt{p_{i\cdot}}} - \frac{p_{ik}}{p_{\cdot k}\sqrt{p_{i\cdot}}}\right)^2} \tag{7-54}$$

其中

$$p_{i\cdot} = \sum_{j=1}^{n} p_{ij} \quad (i = 1,2,\cdots,m)$$

为了计算加权距离，只要把 n 个样品点的坐标改为

$$\left(\frac{p_{1j}}{p_{\cdot j}\sqrt{p_{1\cdot}}}\ \frac{p_{2j}}{p_{\cdot j}\sqrt{p_{2\cdot}}} \cdots \frac{p_{mj}}{p_{\cdot j}\sqrt{p_{m\cdot}}}\right)' \qquad (j = 1,2,\cdots,n) \tag{7-55}$$

即可。

用 P 矩阵中第 i 行元素的和

$$p_{i\cdot} = \sum_{j=1}^{n} p_{ij} \qquad (i = 1,2,\cdots,m)$$

去除以 P 矩阵中第 i 行的各个元素，得

$$\left(\frac{p_{i1}}{p_{i\cdot}}\ \frac{p_{i2}}{p_{i\cdot}}\ \cdots\ \frac{p_{in}}{p_{i\cdot}}\right) \qquad (i = 1,2,\cdots,m) \tag{7-56}$$

式(7-56)表示 n 维空间中 m 个变量点的坐标。两个变量点 l 和 k 之间的加权距离为

$$D^*(l,k) = \sqrt{\sum_{j=1}^{n} \frac{1}{p_{\cdot j}}\left(\frac{p_{lj}}{p_{l\cdot}} - \frac{p_{kj}}{p_{k\cdot}}\right)^2} = \sqrt{\sum_{j=1}^{n}\left(\frac{p_{lj}}{p_{l\cdot}\sqrt{p_{\cdot j}}} - \frac{p_{kj}}{p_{k\cdot}\sqrt{p_{\cdot j}}}\right)^2} \tag{7-57}$$

为了计算变量点之间的加权距离，只需把 m 个变量点的坐标改写为

$$\left(\frac{p_{i1}}{p_{i\cdot}\sqrt{p_{\cdot 1}}}\quad \frac{p_{i2}}{p_{i\cdot}\sqrt{p_{\cdot 2}}}\quad \cdots\quad \frac{p_{in}}{p_{i\cdot}\sqrt{p_{\cdot n}}}\right) \qquad (i = 1,2,\cdots,m) \tag{7-58}$$

便可以了。

为了对样品进行地质解释，现用式(7-55)和式(7-58)进一步研究样品和变量的关系。

二、协方差矩阵和因子载荷矩阵

1. 变量的协方差矩阵

(1) 变量的均值

在矩阵 P 中，若将元素 p_{ij} 视为概率，那么 $p_{i\cdot}$、$p_{\cdot j}$ 就是边缘概率，因此 m 维空间中样品点第 i 个变量的概率均值为

$$\sum_{j=1}^{n} \frac{p_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}} \cdot p_{\cdot j} = \frac{1}{\sqrt{p_{i\cdot}}}\sum_{j=1}^{n} p_{ij} = \sqrt{p_{i\cdot}} \qquad (i = 1,2,\cdots,m)$$

(2) 变量的协方差矩阵

第 i 个变量与第 j 个变量的协方差为

$$\begin{aligned} s_{ij} &= \sum_{k=1}^{n}\left(\frac{p_{ik}}{\sqrt{p_{i\cdot}p_{\cdot k}}} - \sqrt{p_{i\cdot}}\right)\left(\frac{p_{jk}}{\sqrt{p_{j\cdot}p_{\cdot k}}} - \sqrt{p_{j\cdot}}\right) \cdot p_{\cdot k} \\ &= \sum_{k=1}^{n}\left(\frac{p_{ik} - p_{i\cdot}p_{\cdot k}}{\sqrt{p_{i\cdot}p_{\cdot k}}}\right)\left(\frac{p_{jk} - p_{j\cdot}p_{\cdot k}}{\sqrt{p_{j\cdot}p_{\cdot k}}}\right) = \sum_{k=1}^{n} Z_{ik} \cdot Z_{jk} \end{aligned}$$

其中

$$\begin{aligned} Z_{ik} &= \frac{p_{ik} - p_{i\cdot}p_{\cdot k}}{\sqrt{p_{i\cdot}p_{\cdot k}}} = \frac{\left(\frac{x_{ik}}{T} - \frac{x_{i\cdot}}{T}\frac{x_{\cdot k}}{T}\right)}{\left(\frac{x_{i\cdot}}{T}\cdot\frac{x_{\cdot k}}{T}\right)^{1/2}} \\ &= (x_{ik} - x_{i\cdot}x_{\cdot k}/T)(x_{i\cdot}x_{\cdot k})^{-1/2} \end{aligned}$$

如果记

$$S = [s_{ij}] \qquad (i = 1,2,\cdots,m; j = 1,2,\cdots,m)$$
$$Z = [z_{ik}] \qquad (i = 1,2,\cdots,m; k = 1,2,\cdots,n)$$

那么 m 个变量的协方差矩阵

$$S = ZZ' \tag{7-59}$$

2. 样品的协方差矩阵

(1) 样品的概率均值

在 m 维空间中，第 k 个样品的概率均值为

$$\sum_{i=1}^{m} \frac{p_{ik}}{p_{i.}\sqrt{p_{.k}}} \cdot p_{i.} = \frac{1}{\sqrt{p_{.k}}}\sum_{i=1}^{m} p_{ik} = \sqrt{P_{.k}}$$

(2) 样品的协方差

任意两个样品 l 和 k 的协方差为

$$S_{kl}^{*} = \sum_{i=1}^{m}\left[\frac{p_{ik}}{p_{i.}\sqrt{p_{.k}}} - \sqrt{p_{.k}}\right]\left[\frac{p_{il}}{p_{i.}\sqrt{p_{.l}}} - \sqrt{p_{.l}}\right] \cdot p_{i.}$$

$$= \sum_{i=1}^{m}\left[\frac{p_{ik} - p_{i.}p_{.k}}{\sqrt{p_{i.} \cdot p_{.k}}}\right]\left[\frac{p_{il} - p_{i.}p_{.l}}{\sqrt{p_{i.} \cdot p_{.l}}}\right] = \sum_{i=1}^{m} Z_{ik} \cdot Z_{il}$$

其中

$$Z_{ik} = (x_{ik} - x_{i.}x_{.k}/T)(x_{i.}x_{.k})^{-½}$$

记

$$S^{*} = [s_{kl}^{*}] \quad (k = 1,2,\cdots,n; l = 1,2,\cdots,n)$$

从而得样品的协方差矩阵

$$S^{*} = Z'Z \tag{7-60}$$

3. 因子载荷矩阵

由线性代数可知，矩阵 ZZ' 和 $Z'Z$ 有相同的非零特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p (p \leqslant m)$，并且，对其中的每个 $\lambda_j (1 \leqslant j \leqslant p)$，若对应的 u_j 是 ZZ' 的单位特征向量，那么

$$v_j = Z'u_j \tag{7-61}$$

是 $Z'Z$ 相应的单位特征向量；反之，若 v_j 是与 $Z'Z$ 的特征值 λ_j 对应的单位特征向量，那么

$$u_j = Zv_j \tag{7-62}$$

则是 ZZ' 所相应的单位特征向量。

上述结果表明，当求得变量协方差矩阵 S 的特征值 $\lambda_j (j=1,2,\cdots,p)$ 和与其对应的特征向量 $u_j (j=1,2,\cdots,p)$ 后，便可得到 R 型因子分析的因子载荷矩阵，再由式(7-61)可直接求得 Q 型因子分析的因子载荷矩阵，这样就克服了由于样品数量过多带来的 Q 型因子分析在计算上的困难。此外，S 与 S^{*} 有相同的特征值，这些特征值表示各公因子所提供的方差，因而在变量空间中的第一因子、第二因子、……、直到第 p 因子与样品空间中相应的各个因子在总方差中所占的百分比完全相同，因此，又可用相同的因子轴同时表示变量和样品，这样便把 R 型与 Q 型因子分析统一起来。

(1) R 型因子载荷矩阵

若取 S 的前 p 个特征值 $\lambda_1, \lambda_2, \cdots, \lambda_p$，与它们对应的 p 个单位特征向量 $u_1, u_2, \cdots, u_p$，那么因子载荷矩阵为

$$U = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1p}\sqrt{\lambda_p} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2p}\sqrt{\lambda_p} \\ \cdots & \cdots & \cdots & \cdots \\ u_{m1}\sqrt{\lambda_1} & u_{m2}\sqrt{\lambda_2} & \cdots & u_{mp}\sqrt{\lambda_p} \end{bmatrix} \tag{7-63}$$

(2) Q 型因子载荷矩阵

根据式(7-61)求得 Q 型因子载荷矩阵

$$V = \begin{pmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1p}\sqrt{\lambda_p} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2p}\sqrt{\lambda_p} \\ \cdots & \cdots & \cdots & \cdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_2} & \cdots & v_{np}\sqrt{\lambda_p} \end{pmatrix} \tag{7-64}$$

三、对应分析计算步骤

1. 求 Z 矩阵

按式(7-65)把原始数据矩阵 X 变换为 Z 矩阵

$$z_{ij} = \frac{x_{ij} - x_{i.}x_{.j}/T}{\sqrt{x_{i.}x_{.j}}} \qquad (i = 1,2,\cdots,m; j = 1,2,\cdots,n) \tag{7-65}$$

式中

$$x_{i.} = \sum_{j=1}^{n} x_{ij} \qquad (i = 1,2,\cdots,m)$$

$$x_{.j} = \sum_{i=1}^{n} x_{ij} \qquad (j = 1,2,\cdots,n)$$

$$T = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \sum_{i=1}^{m} x_{i.}$$

2. R 型因子分析

(1) 求因子载荷矩阵

求变量协方差矩阵 $S = ZZ'$ 的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$，按其累积百分比 $\left(\sum_{i=1}^{p}\lambda_i \Big/ \sum_{i=1}^{m}\lambda_i\right) \geqslant 85\%$ 取前 p 个特征值 $\lambda_1, \lambda_2, \cdots, \lambda_p$，并计算与它们对应的单位特征向量 $u_1, u_2, \cdots, u_p$，得 R 型因子载荷矩阵式(7-63)。

(2) 在因子平面上作变量的散点图

以因子载荷为坐标、因子作坐标轴把变量点在因子轴平面上表示出来。

3. Q 型因子分析

(1) 求 Q 型因子载荷矩阵

在 R 型因子载荷的基础上，根据式(7-61)求出 Q 型因子载荷式(7-64)。

(2) 在因子平面上作样品散点图

在 R 型因子分析的因子平面上，以 Q 型因子载荷为坐标把样品点表示出来。

§8 因子分析 FORTRAN 源程序

一、因子分析程序

本程序主要用于对具有 m 个变量的 n 个样品进行 R 型或 Q 型因子分析，给出主因子载荷矩阵、方差最大因子载荷矩阵、主因子得分矩阵和方差最大因子得分矩阵。

1. 符号说明

x ——存放原始数据的 n 行 m 列的数组名；

c ——相似性度量矩阵；

v ——存放特征向量的二维数组名；

a1 ——存放特征值的一维数组名；

a ——存放因子载荷的二维数组名；

f ——存放因子得分的二维组名；

t ——正交变换矩阵；

n1 ——样品数，整型变量；

m1 ——变量数，整型变量；

l ——主因子数，整型变量；

ns ——数据预处理方法选择变量；

nd ——相似性度量选择变量；

q —— Q 型因子分析选择变量；

r —— R 型因子分析选择变量。

2. 子程序

stand ——数据标准差标准化子程序；

norm ——数据极差标准化子程序；

rcoef ——计算相关矩阵子程序；

ctheta ——计算相似矩阵子程序；

jacobi —— JACOBI 法求对称矩阵特征值和特征向量子程序；

fczdxz ——方差最大正交旋转子程序；

yzdf ——计算因子得分子程序。

3. 程序使用说明

(1) 数据文件

n 个样品 m 个变量的观测值以数据文件形式存放，其存放格式为：

$$x_{ij} \quad (i = 1,2,\cdots,n; j = 1,2,\cdots,m)$$

(2) 操作说明

在 DOS 操作系统下，键入因子分析目的程序名 yzfx 后的具体操作如下：

①键盘输入样品数 n1 变量数 m1(Input n1,m1)。

②键盘输入数据文件名(Input your data file name).

③选择数据预处理方法；ns=0，数据不作预处理；ns=1，数据标准差标准化；ns=2，数据极差标准化。

④ 选择因子分析类型(To select the mode Q or R)，键入 Q 作 Q 型因子分析，R 则作 R 型因子分析。

⑤ 选择相似性度量：nd=1 为相关系数，nd=2 为相似系数。

⑥ 根据特征值累积百分比。确定主因子数 l。

⑦ 最后，当一次因子分析进行完后，将先后显示：

another similarity metric(y/n)?

another mode(y/n)?

等提示，可根据需要对其回答。

(3) 主要输出结果。

主要结果包括以下几个数据文件：

ru.dat、ra.dat、rdf.dat、rda.dat、rddf.dat 分别为 R 型因子分析的特征向量，因子载荷、因子得分和方差最大因子载荷及因子得分数据文件。

qu.dat、qa.dat、qdf.dat、qda.dat、qddf.dat 分别为 Q 型因子分析的结果。

4. 源程序

(1) 因子分析程序流程

因子分析程序流程如图 7-2 所示。

(2) 因子分析 FORTRAN 源程序。

```
$large
$debug
          parameter(i0=150,i1=30)
          dimension lt(i0),xk(i1),sk(i1)
          dimension a(i0,i0),t(i0,i0),d(i0)
          common /df/x(150,30),f(150,30)
          common /vc/v(150,150),c(150,150),d1(150)
          equivalence (a(1,1),c(1,1)),(t(1,1),v(1,1))
          character fname*10,type,mode
          character *10 fm1,fm2,fm3,fm4,fm5
          write(*,*) ' Factor analysis procedure '
          write(*,*) ' Input n1,m1 '
          read(*,*) n1,m1
          write(*,*) ' Input your data file name '
          read(*,'(a)') fname
100       write(*,*) ' To select a data pre-processing mode '
          write(*,*) ' ns:0,1,2, ns=? '
          read(*,*) ns
          nend=0
          open(1,file=fname)
          do 105 i=1,2000
          read(1,*,err=110,end=115) (x(i,j),j=1,m1)
          nend=nend+1
105       continue
110       write(*,'(a,i4)') 'Error of data file: row=',i
          stop
115       write(*,'(a,i4)') ' End of data file: row=',nend
          close(1)
          if(ns.eq.1) call stand(n1,m1)
          if(ns.eq.2) call norm(n1,m1)
120       write(*,*) ' To select the mode:Q or R '
```

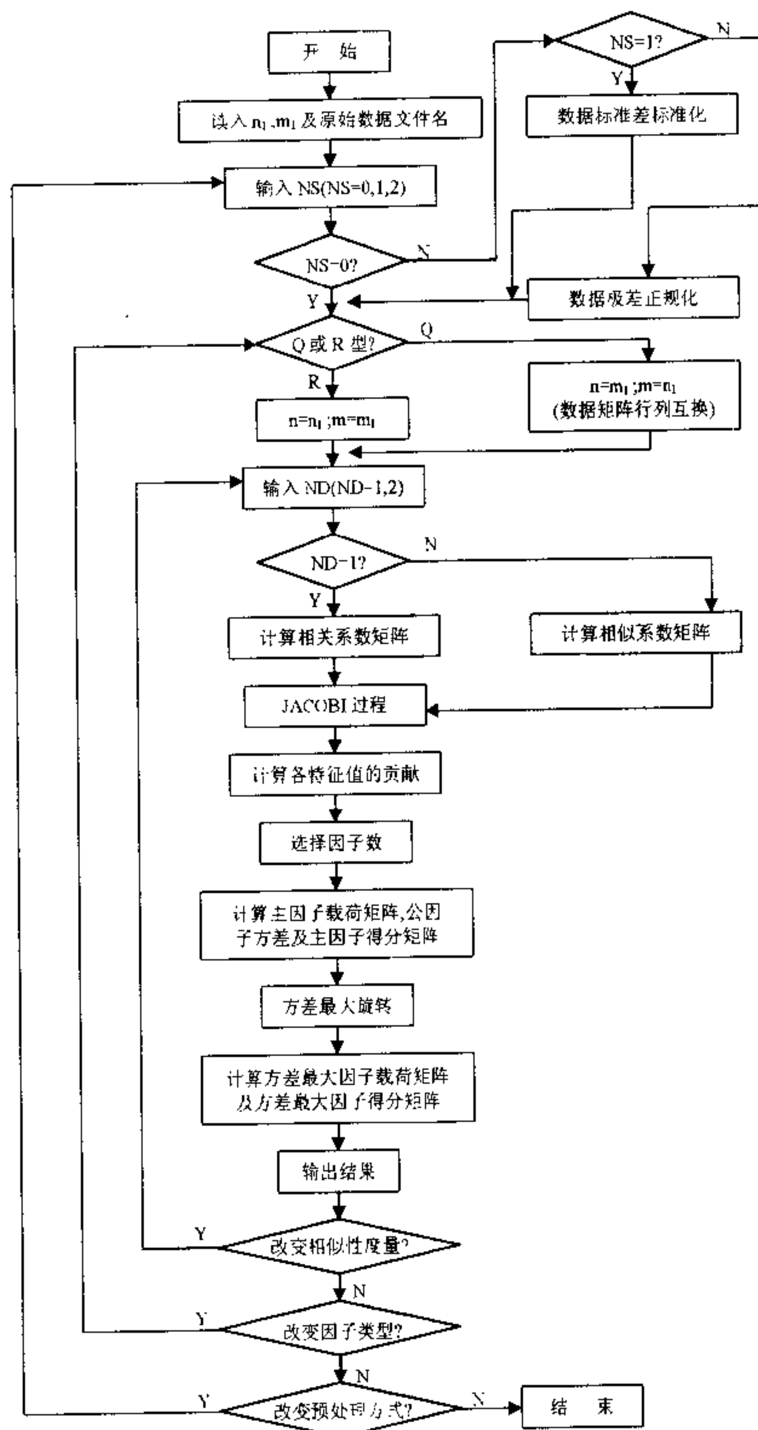


图 7-2 因子分析程序流程图

```
      read( * ,'(a)') mode
      if(mode.eq.'Q'.or.mode.eq.'q') then
      do 130 i=1,n1
      do 125 j=1,m1.
125   v(j,i)=x(i,j)
130   continue
      n=m1
      m=n1
      else
      do 140 i=1,n1
      do 135 j=1,m1
135   v(i,j)=x(i,j)
140   continue
      n=n1
      m=m1
      end if
145   write( * , * ) ' To select a similarity metric '
      write( * , * ) ' nd: 1, 2, nd= '
      read( * , * ) nd
      if(nd.eq.1) call rcoef(n,m)
      if(nd.eq.2) call ctheta(n,m)
      call jacobi(m)
      ds=0.
      do 150 i=1,m
150   lt(i)=i
      do 160 i=1,m-1
      dmax=d1(i)
      kk=i
      do 155 k=i,m
      if(d1(k).gt.dmax) then
      dmax=d1(k)
      kk=k
      end if
155   continue
      dd=d1(i)
      d1(i)=d1(kk)
      d1(kk)=dd
      llt=lt(i)
      lt(i)=lt(kk)
      lt(kk)=llt
```

```
          d(i)=dmax
          ds=ds+d(i)
160       continue
          do 165 i=1,m
165       d1(i)=d(i)/ds * 100.
          write( * ,170)
170       format(/8x,' Eigenvalue percent ')
          ds1=0.
          do 175 i=1,m
          ds1=ds1+d1(i)
175       write( * ,180) d(i),ds1
180       format(7x,f10. 4,5x,f6. 2)
          write( * ,'(a,f8. 4)') ' sum=',ds
          write( * , * ) ' Toselect the number of factor:L=? '
          read( * , * ) l
          if(mode. eq. 'R'. or. mode. eq. 'r') then
          fm1='ru. dat'
          fm2='ra. dat'
          fm3='rdf. dat'
          fm4='rda. dat'
          fm5='rddf. dat'
          open(2,file=fm1)
          open(3,file=fm2)
          open(4,file=fm3)
          open(5,file=fm4)
          open(6,file=fm5)
          else
          fm1='qu. dat'
          fm2='qa. dat'
          fm3='qdf. dat'
          fm4='qda. dat'
          fm5='qddf. dat'
          open(2,file=fm1)
          open(3,file=fm2)
          open(4,file=fm3)
          open(5,file=fm4)
          open(6,file=fm5)
          end if
          do 185 i=1,l
185       write(2,190) i,(v(j,lt(i)),j=1,m)
  190
```

```
190     format(1x,' vector',i3,(t12,10f8.4))
        do 200 i=1,m
        ds1=0.
        do 195 j=1,l
        a(i,j)=v(i,lt(j))*sqrt(d(j))
195     ds1=ds1+a(i,j)**2
200     d1(i)=ds1
        write(*,*)' The principal component factor matrix '
        do 205 i=1,m
205     write(3,210) i,(a(i,j),j=1,l),d1(i)
210     format(1x,i3,',',f8.4,',',f8.4,',',f8.4,',',f8.4,',',f8.4)
        call yzdf(n,m,l,1,mode,d)
        do 211 i=1,n
211     write(4,210) i,(f(i,j),j=1,l)
        call fczdxz(m,l)
        do 225 i=1,m
        xk(i)=0.
225     sk(i)=0.
        ds1=0.
        do 235 i=1,m
        do 230 j=1,l
        ds1=a(i,j)*a(i,j)
        xk(j)=xk(j)+ds1
        sk(i)=sk(i)+ds1
230     continue
235     continue
        do 240 i=1,l
240     d1(i)=xk(i)/ds
        write(*,*)' variance percent '
        ds1=0.
        do 245 i=1,l
        ds1=ds1+d1(i)*100
245     write(*,180) xk(i),ds1
        write(*,*)' Communalities '
        do 260 i=1,m
260     write(5,210) i,(a(i,j),j=1,l),sk(i)
        call yzdf(n,m,l,2,mode,d)
        do 265 i=1,n
265     write(6,210) i,(f(i,j),j=1,l)
        close(2)
```

```
        close(3)
        close(4)
        close(5)
        close(6)
        write( * , * ) ' another similarity metric(y/n) '
        read( * ,'(a)') type
        if(type.eq.'y') go to 145
        write( * , * ) ' another mode (y/n) ? '
        read( * ,'(a)') type
        if(type.eq.'y') go to 120
        write( * , * ) ' another data pre—processing mode (y/n)? '
        read( * ,'(a)') type
        if(type.eq.'y') go to 100
        stop
        end

        subroutine norm(n,m)
        common /df/x(150,30),p(150,30)
        do 140 j=1,m
        xmax=x(1,j)
        xmin=x(1,j)
        do 100 i=2,n
        if(x(i,j).gt.xmax) xmax=x(i,j)
        if(x(i,j).lt.xmin) xmin=x(i,j)
100     continue
        do 120 i=1,n
120     x(i,j)=(x(i,j)-xmin)/(xmax-xmin)
140     continue
        return
        end

        subroutine stand(n,m)
        common/df/x(150,30),p(150,30)
        do 130 i=1,m
        sx=0.
        sxx=0.
        do 110 j=1,n
        sx=sx+x(j,i)
        sxx=sxx+x(j,i) * * 2
110     continue
```

```
      xm=sx/float(n)
      sd=sqrt((sxx-sx*sx/float(n))/float(n-1))
      do 120 j=1,n
      x(j,i)=(x(j,i)-xm)/sd
120   continue
130   continue
      return
      end

      subroutine rcoef(n,m)
      common/vc/x(150,150),r(150,150),dd1(150)
      an=n
      do 120 i=1,m
      do 110 j=1,m
      sx1=0.
      sx2=0.
      sx1x1=0.
      sx2x2=0.
      sx1x2=0.
      do 100 k=1,n
      sx1=sx1+x(k,i)
      sx2=sx2+x(k,j)
      sx1x1=sx1x1+x(k,i)**2
      sx2x2=sx2x2+x(k,j)**2
      sx1x2=sx1x2+x(k,i)*x(k,j)
100   continue
      rr=sqrt((sx1x1-sx1*sx1/an)*(sx2x2-sx2*sx2/an))
      rr=(sx1x2-sx1*sx2/an)/rr
      r(i,j)=rr
      r(j,i)=rr
110   continue
120   continue
      return
      end

      subroutine ctheta(n,m)
      common /vc/x(150,150),c(150,150),d1(150)
      write(*,*)'ctheta matrix'
      do 120 i=1,m
      do 110 j=i,m
```

```
      di=0.
      dj=0.
      dij=0.
      do 100 k=1,n
      di=di+x(k,i) * x(k,i)
      dj=dj+x(k,j) * x(k,j)
100   dij=dij+x(k,i) * x(k,j)
      c(i,j)=dij/sqrt(di * dj)
      c(j,i)=c(i,j)
110   continue
120   continue
      return
      end

      subroutine fczdxz(m,l)
      common /vc/t(150,150),a(150,150),h(150)
      do 10 i=1,l
      do 10 j=1,l
      if(j.eq.i) t(i,j)=1.
      if(j.ne.i) t(i,j)=0.
10    continue
      v0=0.
      do 20 i=1,m
20    h(i)=sqrt(h(i))
      do 40 i=1,m
      do 40 j=1,l
40    a(i,j)=a(i,j)/h(i)
      do 150 k=0,100
      b2=0.
      b4=0.
      do 80 j=1,l
      d2=0.
      do 60 i=1,m
      d2=d2+a(i,j) * * 2
60    b4=b4+a(i,j) * * 4
80    b2=b2+d2 * * 2
      v1=b4/m-b2/m/m
      if((v1-v0).lt.1.e-7) go to 170
      if((v1-v0).ge.1.e-7) v0=v1
      do 150 np=1,l-1
```

```
      do 150 nq=np+1,l
      c1=0.
      c2=0.
      c3=0.
      c4=0.
      do 100 j=1,m
      u1=a(j,np)**2-a(j,nq)**2
      u2=2*a(j,np)*a(j,nq)
      c1=c1+u1
      c2=c2+u2
      c3=c3+u1*u1-u2*u2
100   c4=2*u1*u2+c4
      u1=c4-2*c1*c2/m
      u2=c3-(c1*c1-c2*c2)/m
      fi=atan2(u1,u2)/4
      c=cos(fi)
      s=sin(fi)
      do 120 j=1,m
      c1=a(j,np)*c+a(j,nq)*s
      a(j,nq)=-a(j,np)*s+a(j,nq)*c
120   a(j,np)=c1
      do 130 j=1,l
      c1=t(j,np)*c+t(j,nq)*s
      t(j,nq)=-t(j,np)*s+t(j,nq)*c
130   t(j,np)=c1
150   continue
170   do 180 i=1,m
      do 180 j=1,l
180   a(i,j)=a(i,j)*h(i)
      return
      end

      subroutine yzdf(n,m,l,np,mode,d)
      common /df/z(150,30),f(150,30)
      common /vc/t(150,150),a(150,150),d1(150)
      dimension d(n)
      character mode
      if(np.eq.2) go to 140
      if(mode.eq.'R'.or.mode.eq.'r')then
      do 110 i=1,n
```

```
        do 110 j=1,l
        s=0.
        do 100 k=1,m
100     s=s+z(i,k)*a(k,j)
110     f(i,j)=s/d(j)
        else
        do 130 i=1,n
        do 130 j=1,l
        s=0.
        do 120 k=1,m
120     s=s+z(k,i)*a(k,j)
130     f(i,j)=s/d(j)
        end if
        write(*,*)' the principal factor score matrix'
        go to 210
140     do 200 i=1,n
        do 180 j=1,l
        s=0.
        do 160 k=1,l
160     s=f(i,k)*t(k,j)+s
180     a(i,j)=s
        do 200 j=1,l
200     f(i,j)=a(i,j)
210     continue
        return
        end

        subroutine jacobi(n)
        dimension b(150),z(150)
        common /vc/v(150,150),x(150,150),d(150)
        integer step
        do 120 i=1,n
        do 120 j=1,n
        if(i.eq.j) v(i,j)=1.
        if(i.ne.j) v(i,j)=0.
120     continue
        do 130 i=1,n
        d(i)=x(i,i)
        b(i)=d(i)
130     z(i)=0.
```

```
      do 240 step=1,100
      sm=0.
      do 140 k1=1,n-1
      do 140 k2=k1+1,n
140   sm=sm+abs(x(k1,k2))
      if(sm.eq.0.) go to 250
      if(i.lt.4) tr=0.2*sm/n/n
      if(i.ge.4) tr=0.
      do 220 k1=1,n-1
      do 220 k2=k1+1,n
      g=100.*abs(x(k1,k2))
      if(i.gt.4.and.abs(d(k1)+g).eq.abs(d(k1))
     # .and.abs(d(k2)+g).eq.abs(d(k2))) then
      x(k1,k2)=0.
      else if(abs(x(k1,k2)).gt.tr) then
      h=d(k2)-d(k1)
      if((abs(h)+g).eq.abs(h)) then
      t=x(k1,k2)/h
      else
      th=0.5*h/x(k1,k2)
      t=1./(abs(th)+sqrt(1.+th*th))
      if(th.lt.0.) t=-t
      end if
      c=1./sqrt(1.+t*t)
      s=t*c
      ta=s/(1.+c)
      h=t*x(k1,k2)
      z(k1)=z(k1)-h
      z(k2)=z(k2)+h
      d(k1)=d(k1)-h
      d(k2)=d(k2)+h
      x(k1,k2)=0.
      j1=k1-1
      if(j1.le.0) go to 160
      do 150 j=1,j1
      g=x(j,k1)
      h=x(j,k2)
      x(j,k1)=g-s*(h+g*ta)
150   x(j,k2)=h+s*(g-h*ta)
160   j2=k2-1
```

```
          if(j2.lt.k1+1) go to 180
          do 170 j=k1+1,j2
          g=x(k1,j)
          h=x(j,k2)
          x(k1,j)=g-s*(h+g*ta)
170       x(j,k2)=h+s*(g-h*ta)
180       j1=k2+1
          if(n.lt.j1) go to 200
          do 190 j=j1,n
          g=x(k1,j)
          h=x(k2,j)
          x(k1,j)=g-s*(h+g*ta)
190       x(k2,j)=h+s*(g-h*ta)
200       do 210 j=1,n
          g=v(j,k1)
          h=v(j,k2)
          v(j,k1)=g-s*(h+g*ta)
210       v(j,k2)=h+s*(g-h*ta)
          end if
220       continue
          do 230 k1=1,n
          b(k1)=b(k1)+z(k1)
          d(k1)=b(k1)
230       z(k1)=0.
240       continue
250       write(*,'(a,i3)') 'step=', step
          return
          end
```

二、对应分析程序

本程序用于对具有 m 个变量的 n 个样品进行数据变换，得到 R 型和 Q 型因子分析因子载荷矩阵，选择任二个因子轴绘变量和样品的散点图。

(一) 符号说明

n—样品数，整型变量；

m—变量数，整型变量；

ii—因子数；

x—存放原始数据的二维数组名；

r—存放 R 型因子载荷的二维数组名；

q—存放 Q 型因子载荷的二维数组名。

(二) 子程序

jacobi—用 JACOBI 法求对称矩阵全部特征值和特征向量子程；

mp—绘制因子平面图子程序。

（三）输出结果

R 型和 Q 型因子分析的因子载荷数据文件 y_1.dat、y_2.dat 和因子平面图。

（四）程序使用说明

1. 数据文件

对应分析的数据文件与因子分析中的数据文件相同。

2. 操作说明

在 DOS 操作系统下，键入对应分析目的程序名 dyfx 后的具栖操作如下：

（1）输入数据文件名；

（2）输入样品数 n 和变量数 m；

（3）输入选择的因子数 ii；

（4）输入绘图参数（按屏幕末回答），绘制因子平面图。

（五）源程序

对应分析源程序如下：

```
$large
$debug
          dimension x(200,30),pm(30),pn(200)
          common/oj/a(30,30),s(30,30),d(30)
          common/ot/r(30,30),q(200,30)
          common/oh/y(230,2)
          character *10 fname,fm1,fm2
          common/h/y1,y2
          common/fm/fm1,fm2
          write(*,*)' Enter yourdata file name '
          read(*,'(a)') fname
          write(*,*)' input n,m '
          read(*,*) n,m
          open (1,file=fname)
          n1=0
          do 10 i=1,2000
          read(1,*,err=20,end=30) (x(i,j),j=1,m)
          n1=n1+1
10        continue
20        write(*,'(a,i5)') ' err of file ',i
          stop
30        write(*,'(a,i5)') ' end of file ',n1
          t=0.
          do 50 i=1,n
          pn(i)=0.
          do 40 j=1,m
```

```
40      pn(i)=pn(i)+x(i,j)
50      t=t+pn(i)
        do 60 j=1,m
        pm(j)=0.
        do 60 i=1,n
60      pm(j)=pm(j)+x(i,j)
        do 70 j=1,m
        do 70 i=1,n
70      x(i,j)=(x(i,j)-pm(j) * pn(i)/t)/sqrt(pm(j) * pn(i))
        do 85 i=1,m
        do 85 j=1,i
        t=0.
        do 80 k=1,n
80      t=t+x(k,i) * x(k,j)
        a(i,j)=t
85      a(j,i)=t
        call jacobi(m,1.e-3)
        do 95 i=1,m-1
        do 95 j=i+1,m
        if(d(i).lt.d(j)) then
        d1=d(i)
        d(i)=d(j)
        d(j)=d1
        do 90 k=1,m
        a(k,k)=s(k,i)
        s(k,i)=s(k,j)
90      s(k,j)=a(k,k)
        end if
95      continue
        t=0.
        do 96 i=1,m
96      t=t+d(i)
        write( * , * ) ' eigenvalues percent(%) '
        d1=0.
        do 98 i=1,m
        d1=d(i) * 100./t+d1
        write( * ,'(t3,f8.5,t19,f7.3)') d(i),d1
        if(d1.gt.90.) go to 99
98      continue
99      ii=i
```

```
      write( * ,'(//,a,i2)') ' factor number= ',ii
      do 140 j=1,ii
      d1=sqrt(d(j))
      do 100 i=1,m
100   r(i,j)=d1 * s(i,j)
      do 110 i=1,n
      q(i,j)=0.
      do 110 k=1,m
110   q(i,j)=q(i,j)+x(i,k) * s(k,j)
      t=0.
      do 120 i=1,n
120   t=t+q(i,j) * q(i,j)
      t=sqrt(t)
      do 130 i=1,n
130   q(i,j)=d1 * q(i,j)/t
140   continue
      fm1='y1.dat'
      open(10,file=fm1)
      do 150 i=1,m
150   write(10,'(1x,i4,10f7.4)') i,(r(i,j),j=1,ii)
      fm2='y2.dat'
      open(12,file=fm2)
      do 160 i=1,n
160   write(12,'(1x,i4,10f7.4)') i,(q(i,j),j=1,ii)
      close(10)
      close(12)
      call mp
      stop
      end

      subroutine jacobi(n,eps)
      common/oj/x(30,30),s(30,30),d(30)
      integer step
      do 10 i=1,n
      do 10 j=1,i
      if(i.eq.j) then
      s(i,j)=1.
      else
      s(i,j)=0.
      s(j,i)=0.
```

```
        end if
10      continue
        t=0.
        do 20 i=2,n
        do 20 j=1.i-1
20      t=t+2*x(i,j)*x(i,j)
        z1=sqrt(t)
        z2=eps/n*z1
        thr=z1
        ind=0
        do 70 step=1,1000
        thr=thr/n
        do 60 k=1,1000
        do 50 nq=2,n
        do 50 np=1,nq-1
        if(abs(x(np,nq)).ge.thr) then
        ind=1
        v1=x(np,np)
        v2=x(np,nq)
        v3=x(nq,nq)
        u=0.5*(v1-v3)
        if(u.eq.0.) then
        ga=-1.
        else
        ga=-1*sign(1.,u)*v2/sqrt(v2*v2+u*u)
        end if
        st=ga/sqrt(2*(1+sqrt(1-ga*ga)))
        ct=sqrt(1-st*st)
        do 30 i=1,n
        t=x(i,np)*ct-x(i,nq)*st
        x(i,nq)=x(i,np)*st+x(i,nq)*ct
        x(i,np)=t
        t=s(i,np)*ct-s(i,nq)*st
        s(i,nq)=s(i,np)*st+s(i,nq)*ct
30      s(i,np)=t
        do 40 i=1,n
        x(np,i)=x(i,np)
40      x(nq,i)=x(i,nq)
        x(np,np)=v1*ct*ct+v3*st*st-2*v2*st*ct
        x(nq,nq)=v1*st*st+v3*ct*ct+2*v2*st*ct
```

```
      x(np,nq)=(v1-v3)*st*ct+v2*(ct*ct-st*st)
      x(nq,np)=x(np,nq)
      end if
50    continue
      if(ind.eq.1) then
      ind=0
      go to 60
      else if(thr.gt.z2) then
      go to 70
      else
      go to 80
      end if
60    continue
70    continue
80    do 90 i=1,n
90    d(i)=x(i,i)
      return
      end

      subroutine mp
      dimension xq(3000),yq(3000),xr(100),yr(100),mx(3000),my(100)
      character*10 fm1,fm2
      common/fm/fm1,fm2
      open(1,file=fm2)
      open(2,file=fm1)
      write(*,'(a))') 'Enter factor total axises:'
      read(*,*) mma
      write(*,'(a))') 'Enter Fn(N) and Fm(M) :'
      read(*,*) n,m
      if(m.le.n) stop 'M<=N,Error!'
      write(*,'(a))') 'Change data? (Y/N)'
      read(*,'(a)') yn
      if(yn.eq.'y'.or.yn.eq.'Y') then
      write(*,'(a))') 'Enter factor(sc):'
      read(*,'(a)') sc
      end if
      write(*,'(a))') 'Enter map position of start(SP) and end(EP):'
      read(*,*) sp,ep
      write(*,'(a))') 'Enter Q-lable type(0,1,...,17):'
      read(*,*) itypeq
```

```
      write( * ,'(a )') ' Enter R—lable type(0,1,. . . ,17): '
      read( * , * ) ityper
      call in
      call fact(10. 0)
      call pen(2)
      call setsty('set10. sym')
      zzzz=9999.
      dx=0. 2
      ntxy=6
      n1=0
      do 10 i=1,3000
      read(1, * ,end=15) mx(i),(xxx,k=1,n-1),xq(i),(xxx,k=n+1,m-1)
     # ,yq(i),(xxx,k=m+1,mma)
      n1=n1+1
10    continue
15    write( * ,55) n1
55    format(1x,' Lines of Qmode—File: ',i4)
      n2=0
      do 20 i=1,100
      read(2, * ,end=25) my(i),(yyy,k=1,n-1),xr(i),(yyy,k=n+1,m-1)
     # ,yr(i),(xxx,k=m+1,mma)
      n2=n2+1
20    continue
25    write( * ,66) n2
66    format(1x,' Lines of Rmode—File: ',i4)
      ees=(ep+sp) * 0. 5
      call movea(sp,ees)
      call linea(ep,ees)
      call movea(ees,sp)
      call linea(ees,ep)
      call text(ep+0. 5 * dx,ees,dx,0. ,'f')
      call numbl1(ep-2. 0 * dx,ees,dx,0. ,n)
      call text(ees-0. 3 * dx,ep+0. 6 * dx,dx,0. ,'f')
      call numbl1(ees-3. 0 * dx,ep+0. 6 * dx,dx,0. ,m)
      call absmax(xq,yq,n1,qmax)
      call absmax(xr,yr,n2,rmax)
      d21=amax0(qmax,rmax)
      if(yn. eq. 'y'. or. yn. eq. 'Y') then
      call flbh(xq,yq,n1,sc,d21)
      call flbh(xr,yr,n2,sc,d21)
```

```
        call absmax(xq,yq,n1,qmax)
        call absmax(xr,yr,n2,rmax)
        d21=amax0(qmax,qmax)
        end if
        if(zzzz.eq.9999.0) then
        dposi=(ep-sp)/4.0
        dtic=(ep-sp)/(2*ntxy)
        do 30 xy=sp,ep+0.05,dtic
        call movea(xy,ees)
        call linea(xy,ees+0.1)
        call movea(ees,xy)
        call linea(ees+0.1,xy)
30      continue
        call numbl2(ees-dposi-5.0*dx,ees+0.2,dx,0.0,-0.5*d21)
        call numbl2(ees-2.0*dx,ees-dposi-0.5*dx,dx,0.0,-0.5*d21)
        call numbl2(ees+dposi-5.0*dx,ees+0.2,dx,0.0,0.5*d21)
        call numbl2(ees-2.0*dx,ees+dposi-0.5*dx,dx,0.0,0.5*d21)
        end if
        des=(ep-sp)*0.5
        call pen(14)
        do 40 i=1,n1
        xx=ees+(xq(i)/d21)*des
        yy=ees+(yq(i)/d21)*des
        call post(xx,yy,0.2,itypeq)
        call setsty("set10.sym")
        call numbl1(xx-4.0*dx,yy-2.0*dx,dx,0.,mx(i))
40      continue
        call pen(12)
        do 50 i=1,n2
        xx=ees+(xr(i)/d21)*des
        yy=ees+(yr(i)/d21)*des
        call post(xx,yy,0.2,ityper)
        call setsty("set10.sym")
        call numbl1(xx-4.0*dx,yy-2.0*dx,dx,0.,my(i))
50      continue
        end

        subroutine absmax(x,y,n,dmax)
        dimension x(n),y(n)
        ax=abs(x(1))
```

```
      ay=abs(y(1))
      do 25 i=1,n
      if(ax.lt.abs(x(i))) ax=abs(x(i))
      if(ay.lt.abs(y(i))) ay=abs(y(i))
25    continue
      dmax=amax0(ax,ay)
      end

      subroutine flbh(x,y,n,sc,dmax)
      dimension x(n),y(n)
      do 10 i=1,n
      bbb=abs(x(i))
      ccc=abs(y(i))
      ddd=dmax*(1.0-sc**bbb)
      eee=dmax*(1.0-sc**ccc)
      x(i)=sign(ddd,x(i))
      y(i)=sign(eee,y(i))
10    continue
      end

      subroutine post(x0,y0,hl,ik)
      real xx(200),yy(200)
      h=0.5*hl
      if(ik.lt.0.or.ik.gt.17) ik=0
      if(ik.eq.0.or.ik.eq.1) then
      call movea(x0-h,y0)
      call linea(x0+h,y0)
      call movea(x0,y0-h)
      call linea(x0,y0+h)
      if(ik.eq.1) then
      call movea(x0+0.707*h,y0-0.707*h)
      call linea(x0-0.707*h,y0+0.707*h)
      call movea(x0-0.707*h,y0-0.707*h)
      call linea(x0+0.707*h,y0+0.707*h)
      end if
      end if
      if(ik.eq.2.or.ik.eq.3) then
      call cir(x0,y0,h,20)
      if(ik.eq.3) then
      if(h.gt.0.02) then
```

```
          do 10 ah=h,0.02,-0.02
          call cir(x0,y0,ah,20)
10        continue
          end if
          end if
          end if
          if(ik.eq.4.or.ik.eq.5) then
          call movea(x0-h,y0-h)
          call linea(x0+h,y0-h)
          call linea(x0+h,y0+h)
          call linea(x0-h,y0+h)
          call linea(x0-h,y0-h)
          if(ik.eq.5) then
          if(h.gt.0.02) then
          do 20 ah=x0-h,x0+h,0.02
          call movea(ah,y0-h)
          call linea(ah,y0+h)
20        continue
          end if
          end if
          end if
          if(ik.eq.6.or.ik.eq.7) then
          aj=210.0/180.0*3.1416
          aj1=330.0/180.0*3.1416
          call movea(x0,y0+h)
          call linea(x0+h*cos(aj),y0+h*sin(aj))
          call linea(x0+h*cos(aj1),y0+h*sin(aj1))
          call linea(x0,y0+h)
          if(ik.eq.7) then
          if(h.gt.0.02) then
          do 30 ah=h,0.02,-0.02
          call movea(x0,y0+ah)
          call linea(x0+ah*cos(aj),y0+ah*sin(aj))
          call linea(x0+ah*cos(aj1),y0+ah*sin(aj1))
          call linea(x0,y0+ah)
30        continue
          end if
          end if
          end if
          if(ik.ge.8) then
```

```
              if(ik.eq.8.or.ik.eq.9) jiaos=4
              if(ik.eq.10.or.ik.eq.11) jiaos=5
              if(ik.eq.12.or.ik.eq.13) jiaos=6
              if(ik.eq.14.or.ik.eq.15) jiaos=7
              if(ik.eq.16.or.ik.eq.17) jiaos=8
              sc=0.4
              djiao=2.0*3.1416/jiaos
              astart=0.5*3.1416
              aend=2.5*3.1416+0.1
              n=0
              do 40 ai=astart,aend,djiao
              n=n+1
              xx(2*n-1)=x0+h*cos(ai)
              yy(2*n-1)=y0+h*sin(ai)
              xx(2*n)=x0+sc*h*cos(ai+0.5*djiao)
              yy(2*n)=y0+sc*h*sin(ai+0.5*djiao)
40            ontinue
              call movea(xx(1),yy(1))
              do 50 i=2,2*n
              call linea(xx(i),yy(i))
50            continue
              if(mod(ik,2).ne.0) then
              if(h.gt.0.02) then
              do 80 ah=h,0.02,-0.02
              n=0
              do 60 ai=astart,aend,djiao
              n=n+1
              xx(2*n-1)=x0+ah*cos(ai)
              yy(2*n-1)=y0+ah*sin(ai)
              xx(2*n)=x0+sc*ah*cos(ai+0.5*djiao)
              yy(2*n)=y0+sc*ah*sin(ai+0.5*djiao)
60            continue
              call movea(xx(1),yy(1))
              do 70 i=2,2*n
              call linea(xx(i),yy(i))
70            continue
80            continue
              end if
              end if
              end if
```

```
      end

      subroutine cir(x0,y0,r,n)
      dt=2.0*3.1415926/n
      call movea(x0+r,y0)
      do 10 i=2,n+1
      t=(i-1)*dt
      x=x0+r*cos(t)
      y=y0+r*sin(t)
      call linea(x,y)
10    continue
      end
```

§9 应用算例

【例1】 盐泉分类与成因

云南某地20个盐泉水化学分析数据见表7-3。为了对盐泉进行分类和解释它们的成因，现对20个样品进行对应分析。

表 7-3 盐泉水化学分析数据

样品序号	矿化度 /g/l	$\frac{Br\cdot10^3}{Cl}$	$\frac{K\cdot10^3}{\sum 盐}$	$\frac{K\cdot10^3}{Cl}$	$\frac{Na}{K}$	$\frac{Mg\cdot10^3}{Cl}$	$\frac{\epsilon Na}{\epsilon Cl}$
1	11.8530	0.4800	14.3600	25.2100	25.2100	0.8100	0.9800
2	45.5960	0.5260	13.8500	24.0400	26.0100	0.9100	0.9600
3	3.5250	0.0860	24.4000	49.3000	11.3000	6.8200	0.8500
4	3.6810	0.3700	13.5700	25.1200	26.0000	0.8200	1.0100
5	48.2870	0.3860	14.5000	25.9000	23.3200	2.1800	0.9300
6	17.9560	0.2800	9.7500	17.0500	37.2000	0.4640	0.9800
7	7.3700	0.5060	13.6000	34.2100	10.6900	8.8000	0.5600
8	4.2230	0.3400	3.8000	7.1000	88.2000	1.1100	0.9700
9	6.4420	0.1900	4.7000	9.1000	23.2000	0.7400	1.0800
10	16.2340	0.3900	3.4000	5.4000	121.5000	0.4200	1.0000
11	10.5850	0.4200	2.4000	4.7000	135.6000	0.8700	0.9800
12	23.5350	0.2300	2.6000	4.6000	141.8000	0.3100	1.0200
13	5.3980	0.1200	2.8000	6.2000	111.2000	1.1400	1.0700
14	283.1480	0.1480	1.7630	2.9680	215.8600	0.1400	0.9800
15	316.6040	0.3170	1.4530	2.4320	263.4099	0.2490	0.9800
16	307.3101	0.1730	1.6270	2.7290	235.7000	0.2140	0.9900
17	322.5149	0.3120	1.3820	2.3200	282.2100	0.0240	1.0000
18	256.5801	0.2970	0.8990	1.4760	410.3000	0.2390	0.9300
19	304.0920	0.2830	0.7890	1.3570	438.3601	0.1930	1.0100
20	240.4460	0.0420	0.7410	1.2660	500.7700	0.2900	0.9900

矩阵 ZZ' 的前两个特征值 λ_1,λ_2 所代表的方差已占总方差的95.94%，因此，取两个主因子即可较好地反映原始数据的变化。根据 λ_1、λ_2 求得的第1、2主因子载荷见表7-4和表7-5。

表 7-4 *R* 型前两个因子载荷

变量序号	变量	f_1	f_2
1	矿化度/g/l	−1.442	0.2286
2	$(Br/Cl)10^3$	0.0343	−0.0084
3	$(K/\sum盐)10^3$	0.3472	0.0104
4	$(K/CL)10_3$	0.5042	0.0144
5	Na/K	−0.1159	−0.1977
6	$(Mg/CL)10^3$	0.1952	0.0022
7	εNa/εCL	0.0454	−0.0200
特征值		0.4503	0.0922
特征值累积百分数		79.65	95.94

表 7-5 Q 型前两个因子载荷

样品序号	f_1	f_2	样品序号	f_1	f_2
1	0.2000	−0.0007	11	−0.0051	−0.1129
2	0.1464	0.0488	12	−0.0131	−0.0965
3	0.3848	0.0142	13	0.0119	−0.1060
4	0.2108	−0.0176	14	−0.0760	0.0887
5	0.1610	0.0560	15	−0.0842	0.0814
6	0.1197	−0.0093	16	−0.0807	0.0916
7	0.3076	0.0153	17	−0.0867	0.0746
8	0.0291	−0.0919	18	−0.0884	−0.0271
9	0.0894	−0.0214	19	−0.0953	−0.0100
10	0.0005	−0.0944	20	−0.0921	−0.0737

对应分析第1、2主因子平面图如图7-3所示。

由于第1个因子 f_1 的方差占总方差的79%以上，因此可以说它是区内起主导作用的一个地质因素，基本上能够反映本区沉积环境演变的主要特征。在图7-3中，f_1 轴的左端为钠、右端为钾，含钾盐泉的主要特征变量 $(K/Cl)10^3$、$(K/\sum盐)10^3$、$(Mg/Cl)10^3$、εNa/εCl、$(Br/Cl)10^3$ 都分布在 f_1 轴的右端，严格受 f_1 控制。f_1 因子载荷中所占比重最大的变量是 $(K/Cl)10^3$（0.5042），位于 f_1 轴的最右端，对于其他变量，按其因子载荷在 f_1 中所占比重的大小自右至左依次排列为：$(K/\sum盐)10^3$（0.3472）、$(Mg/Cl)10^3$（0.1952）、εNa/εCl（0.0454）、$(Bt/Cl)10^3$（0.0343）、Na/K（−0.1159）。这表明了各种盐类物质随着沉积环境的变化而开始沉积的先后次

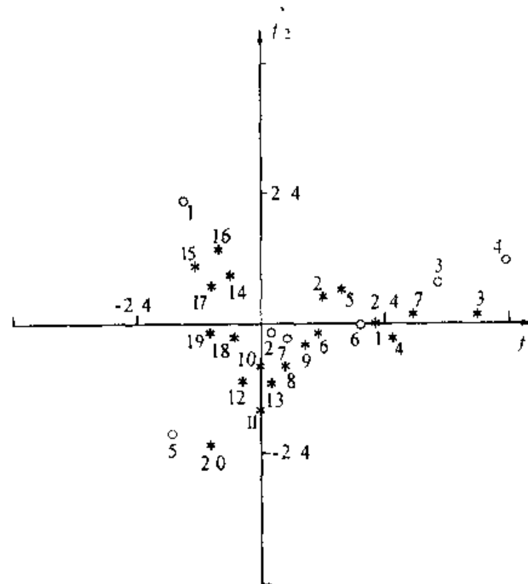


图 7-3 盐泉与水化学对应分析图

○变量 *盐泉

序，钾的浓度自左至右递增，而钠的浓度变化却相反。因此，可以把 f_1 视为盐类沉积的分异轴。钠位于 f_1 轴的左端，反映了富 Na 的沉积环境。$(Br/Cl)10^3$、$(Mg/Cl)10^3$ 位于 f_1 轴靠近坐标轴原点处，说明 Mg 盐与 Br 化物在本区钠盐和钾盐沉积过程中具有一定的浓度，并且钾盐沉积过程中 Mg 盐的混入要比 Br 化物更为明显。

因子 f_2 提供的方差要比 f_1 小的多，仅占总方差的 16.30%。在这一因子中，因子载荷所占比重较大的变量是矿化度(0.2286)和(Na/K)(−0.1977)，分别位于 f_2 轴的上、下两端。这两个变量虽然受因子 f_1 的影响，但更主要的是受 f_2 的支配，而其他变量位于 f_2 轴的中部，在 f_2 所占因子载荷比重都很小，说明它们主要受 f_1 的控制。f_2 因子轴主要反映了含盐地层沉积过程中其他矿物质的混入情况，例如 Fe_2O_3、硫酸盐类、有机盐的混入等等。这表明富钠环境的沉积阶段杂质的混入更为明显。

根据上述分析，按变量与样品分布情况，在图 7-3 上可把样品点划分为三个区。其中，第Ⅰ区位于 f_1 轴右端，样品特征偏于含钾，表明样品区有利于形成钾盐矿床；第Ⅱ区位于 f_1 轴中部，该区主要受 f_2 影响，在靠近 Na/K 方向，盐泉有一定程度的钾矿化，沉积顺序属过渡型，盐泉中有一定浓度的 Na，另外，其他混入物质也相对增加。第Ⅲ区位于 f_1 轴左边，样品特征偏于含钠，该区内样品的 f_1 因子载荷很相近，而 f_2 因子载荷变化范围较大，说明钠盐受矿化度的影响较大。

【例 2】 古潜山油气藏成油地质参数与油气资源量分析

对已探明的一些古潜山油气藏进行分类，研究它们的资源量与成油地质参数的关系，这是评价未知古潜山油气资源的基础。

与其他类型的油气藏一样，油源、供油、储集、圈闭和保存条件是形成古潜山油气藏的基本条件。在此把上述基本条件拟定为 8 项成油地质参数：

x_1 ——古潜山到凹陷生油区中心的距离，km；

x_2 ——储集体的总孔隙度，%；

x_3 ——古潜山上伏盖层的厚度，km；

x_4 ——供油窗口，km^2；

x_5 ——生油岩与不整合面的接触面积，km^2；

x_6 ——闭合高度，km；

x_7 ——闭合面积，km^2；

x_8 ——古潜山最浅埋藏深度，km。

由于存在同一个生油凹陷向不同古潜山圈闭供油气的问题，因此本例未把凹陷的生油量(油源)条件列入成油地质参数。已探明的 26 个古潜山油气藏的成油地质参数见表 7-6。

表 7-6 古潜山成油地质参数表

油藏序号	成油地质参数							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	4.000	3.170	1.260	0.720	13.500	0.700	13.000	2.400
2	12.000	4.670	1.300	0.500	6.500	0.450	6.000	2.300
3	12.000	8.600	0.500	2.600	6.000	0.300	5.200	0.110
4	8.000	10.000	0.800	0.018	4.000	0.300	3.200	1.400
5	8.000	4.140	1.000	0.015	4.000	0.350	4.000	1.700

续表 7-6

油藏序号	成油地质参数							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
6	2.000	3.990	2.350	6.400	10.600	1.200	27.000	2.600
7	3.000	3.860	2.230	0.300	9.400	0.500	9.700	2.400
8	3.000	3.320	1.600	1.440	11.700	0.700	8.400	1.600
9	5.500	3.490	0.560	1.640	5.320	0.350	5.900	0.650
10	3.500	3.700	2.280	5.000	26.600	0.700	26.600	2.500
11	17.500	3.720	2.600	16.000	183.000	1.913	183.000	2.600
12	6.500	2.790	3.968	3.000	4.300	0.350	4.300	3.968
13	2.500	11.030	2.300	10.000	24.000	1.000	24.000	2.300
14	8.000	2.790	2.800	10.000	1.200	0.250	1.200	2.800
15	4.000	2.860	5.050	6.250	4.000	0.450	4.000	5.050
16	3.000	4.160	4.037	1.290	5.500	0.263	5.500	4.037
17	4.000	2.760	4.500	6.750	4.000	0.300	4.000	4.500
18	5.000	4.710	3.850	2.000	6.240	0.550	6.240	3.850
19	8.000	5.820	3.100	9.380	1.840	0.200	1.840	3.100
20	10.000	3.170	1.950	7.500	5.900	0.400	5.900	1.950
21	8.000	1.930	3.150	12.000	3.400	0.350	3.400	3.150
22	20.000	2.500	3.750	6.000	7.000	0.600	7.000	3.750
23	20.000	1.740	4.400	6.000	16.540	0.950	16.500	4.000
24	20.000	1.420	3.700	6.000	4.000	0.650	4.000	3.700
25	8.000	5.500	0.950	5.800	70.000	1.000	64.900	3.500
26	1.000	18.000	0.926	0.150	0.200	0.600	1.800	2.146

根据表 7-6 给出的成油地质参数进行对应分析，计算的特征值、特征值累积百分数及对变分类的第 1、2 公因子载荷见表 7-7。Q 型因子分析的第 1、2 公因子载荷见表 7-8。对应分析第 1、2 公因子平面图如图 7-4 所示。分类结果见表 7-9。

表 7-7　特征值及 R 型第 1、2 公因子载荷

特征值序号	特征值	特征值百分数	特征值累积百分数	变量序号	公因子载荷	
					f_1	f_2
1	0.3168	57.6670	57.6670	1	0.2601	−0.1100
2	0.1298	23.4910	81.1580	2	0.2163	0.3083
3	0.0645	11.6660	92.8240	3	0.1559	−0.0581
4	0.0316	5.7140	98.5380	4	0.1650	0.1352
5	0.0062	1.1290	99.6670	5	−0.2558	0.0026
6	0.0011	0.2010	99.8680	6	0.0272	0.0149
7	0.0007	0.1320	100.0000	7	−0.2519	0.0098
8	0.0000	0.0000	100.0000	8	0.1516	−0.0261

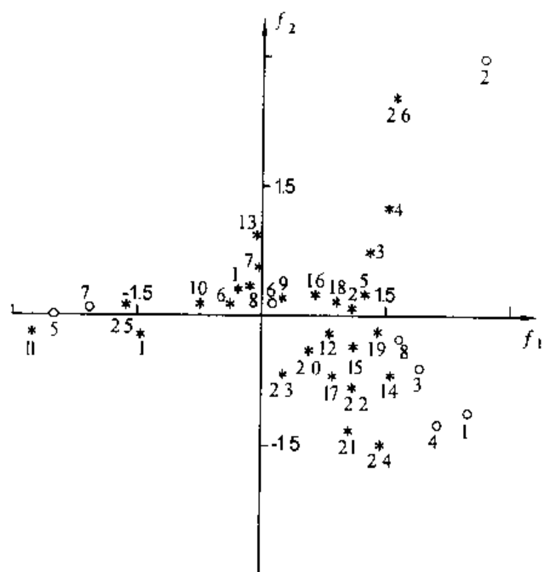


图 7-4 油气藏与成油地质参数对应分析图

○地质参数　　＊油藏

表 7-8 Q 型因子分析第 1、2 公因子载荷

油藏序号	公因子载荷		油藏序号	公因子载荷	
	f_1	f_2		f_1	f_2
1	−0.0296	0.0185	14	0.1387	−0.0680
2	0.0710	0.0090	15	0.0987	−0.0371
3	0.0891	0.0526	16	0.0579	0.0201
4	0.0985	0.1082	17	0.0949	−0.0395
5	0.0662	0.0249	18	0.0643	0.0160
6	−0.0342	0.0039	19	0.1432	−0.0215
7	−0.0026	0.0313	20	0.0832	−0.0434
8	−0.0121	0.0208	21	0.1178	−0.0847
9	0.0294	0.0186	22	0.1177	−0.0737
10	−0.0717	0.0029	23	0.0521	−0.0690
11	0.3222	−0.0186	24	0.1405	−0.0924
12	0.0858	−0.0238	25	0.1689	0.0098
13	−0.0050	0.0741	25	0.1431	0.2636

表 7-9　油气藏分类结果

分类号	原油藏序号	成油地质参数							
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	12	6.500	2.790	3.968	3.000	4.300	0.350	4.300	3.968
	14	8.000	2.790	2.800	10.000	1.200	0.250	1.200	2.800
	15	4.000	2.860	5.050	6.250	4.000	0.450	4.000	5.050
	17	4.000	2.760	4.500	6.750	4.000	0.300	4.000	4.500
	19	8.000	5.820	3.100	9.380	1.840	0.200	1.840	3.100
	20	10.000	3.170	1.950	7.500	5.900	0.400	5.900	1.950
	21	8.000	1.930	3.150	12.000	3.400	0.350	3.400	3.150
	22	20.000	2.500	3.750	6.000	7.000	0.600	7.000	3.750
	23	20.000	1.740	4.400	6.000	16.540	0.950	16.500	4.000
	24	20.000	1.420	3.700	6.000	4.000	0.650	4.000	3.700
2	2	12.000	4.670	1.300	.500	6.500	.450	6.000	2.300
	3	12.000	8.600	.500	2.600	6.000	.300	5.200	.110
	4	8.000	10.000	.800	.018	4.000	.300	3.200	1.400
	5	8.000	4.140	1.000	.015	4.000	.350	4.000	1.700
	9	5.500	3.490	.560	1.640	5.320	.350	5.900	.650
	16	3.000	4.160	4.037	1.290	5.500	.263	5.500	4.037
	18	5.000	4.710	3.850	2.000	6.240	.550	6.240	3.850
	26	1.000	18.000	.926	.150	.200	.600	1.800	2.146
3	1	4.000	3.170	1.260	.720	13.500	.700	13.000	2.400
	6	2.000	3.990	2.350	6.400	10.600	1.200	27.000	2.600
	7	3.000	3.860	2.230	.300	9.400	.500	9.700	2.400
	8	3.000	3.320	1.600	1.440	11.700	.700	8.400	1.600
	10	3.500	3.700	2.280	5.000	26.600	.700	26.600	2.500
	11	17.500	3.720	2.600	16.000	183.000	1.913	183.000	2.600
	13	2.500	14.030	2.300	10.000	24.000	1.000	24.000	2.300
	25	8.000	5.500	.950	5.800	70.000	[illegible]	64.900	3.530
1	平均值	10.850	2.778	3.637	7.288	5.128	.450	5.214	3.579
2		6.813	7.221	1.622	1.027	4.720	.395	4.730	2.024
3		5.438	5.161	1.946	5.707	43.60	.964	44.575	2.491

由特征值累积百分数可知，两个公因子可提取原始数据信息的 81.2%，并且第 1 个公因子 f_1 提供的方差占总方差的 57.7%。由此可见，在形成古潜山油气藏中，它是占主导作用的一个地质因素。在图 7-4 中，反映古潜山油气藏形成条件的主要成油地质参数主要是受 f_1 控制，沿 f_1 轴从左向右的分布顺序依次为 x_5、x_6、x_7、x_8、x_3、x_4 和 x_1。根据成油地质参数的分布特征，可把 f_1 轴理解为构造起伏和断裂作用。成油地质参数 x_2（总孔隙度）主要受 f_2 的控制，可将其主要的地质意义理解为风化剥蚀作用。

在图 7-4 中，26 个古潜山油气藏被明显地划分为三类（见表 7-9），各类的主要特点是：

第 1 类中有 10 个古潜山油气藏，它的主要特点是距生油凹陷中心较远，平均距离为 10.85 公里。另外，古潜山不整合面与生油岩的接触面积较小，主要靠断层供油，埋藏较深，受风化淋滤时间较短，总孔隙度低。由上述可知，该类古潜山圈闭不易捕获大量的油气，探明资源量的平均值为 5.45 百分吨。

第 2 类中有 8 个古潜山油气藏，它们的主要特征是受风化淋滤作用较强，总孔隙度较高。虽然生油岩与古潜山不整合面的接触面积不算大，但占居离生油凹陷中心较近的优越条件，因此这样的古潜山圈闭可以聚集到较多的油气，形成资源较丰富的油气藏，该类油藏探明资源量

的平均值达10百万吨以上。

第3类包括8个古潜山油气藏，它们的显著特点是距生油凹陷中心近，古潜山不整合面与生油岩接触面积大，并且有较高的团合高度和较大的闭合面积。这类古潜山圈闭具备优越的成油地质条件，可以形成含油气丰富的古潜山油气藏，8个油藏探明资源量的平均值超过105百万吨，个别油藏可达500百万吨以上。

综上所述，近油源的不整合面与生油岩接触面积较大的褶皱山和残山易形成含油气丰富的古潜山油气藏。

【例3】 砂岩分类

东濮凹陷的18个砂岩样品分析数据(据赵旭东，修改)见表7-10。以砂岩的三种主要碎屑成分即石英、长石及岩屑的百分含量对砂岩分类。18个样品均有人工鉴定的命名结果，通过对应分析检查分类命名结果是否正确。

表7-10 岩样分析数据及岩石名称

样品号	石英含量/%	长石含量/%	岩屑含量/%	其他含量/%	岩石鉴定命名
1	82	8	9	1	石英砂岩
2	83	8	9	0	石英砂岩
3	86	7	5	2	石英砂岩
4	88	8	12	2	硬砂质石英砂岩
5	78	7	12	3	硬砂质石英砂岩
6	84	5	11	0	硬砂质石英砂岩
7	76	9	13	2	硬砂质石英砂岩
8	61	26	13	0	长石砂岩
9	60	25	15	0	长石砂岩
10	58	29	13	0	长石砂岩
11	57	27	16	0	长石砂岩
12	55	30	15	0	长石砂岩
13	47	27	26	0	混合砂岩
14	46	25	29	0	混合砂岩
15	58	7	35	0	硬砂岩
16	53	17	30	0	长石质硬砂岩
17	61	12	25	2	长石质硬砂岩
18	50	17	33	0	长石质硬砂岩

两个特征值累积百分比为100%，因子平面图如图7-5所示。在图7-5上，18个样品分成四类，与人工分类命名结果完全吻合。这个结果表明，用对应分析进行成因分析是一种有效的方法。

【例4】 沉积有机质类型研究

胜利油田地质科学研究院(1982)根据济阳坳陷渐新世有代表性的14个不成熟生油岩样品和16个泥炭土及湖底、海底现代沉积物样品的有机地化分析数据，利用因子分析方法研究了济阳坳陷的有机质类型。

样品的地质特征参数及选用的地化指标分别见表7-11和表7-12。

对30个样品28项指标进行R型因子分析。方差最大正交旋转后的第1、2因子载荷平面图如图7-6所示。在图7-6上，28项地化指标被划分为三个区：

Ⅰ区的指标是沥青质、氯仿A含量及高碳数部份正烷烃；

Ⅱ区的指标是芳烃及C_{22}含量；

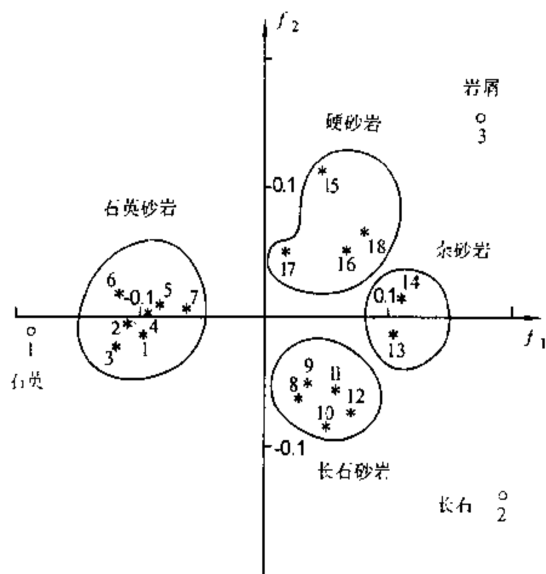


图 7-5　砂岩样品与碎屑因子平面图

Ⅱ区的指标是非烃、氧的含量及低碳数部份正烷烃。

如果把各区特征指标的数据比较一下，见表 7-13，可见到各区的特征指标其数值也是突出的，从而指示了沉积相的特征。

Ⅰ区：植物泥炭土有机质主要指标；

Ⅱ区：海底淤泥及直接受海洋影响的泻湖淤泥类有机质主要指标；

Ⅲ区：内陆淡水湖堰塞湖有机质主要指标。

表 7-11　样品地质特征参数表

样品号	井　号	层　位	井　段/m	岩　性
1	永 31	ES_3 上	1675.00—1680.00	灰色泥岩
2	永 31	ES_3 上	1838.00—1842.00	灰色泥岩
3	永 31	ES_3 上	1857.00　1865.00	灰色泥岩
4	永 31	ES_3 中	1988.00—2024.00	灰色泥岩
5	滨 9	ES_3	1900.00—2090.00	灰色泥岩
6	通 23	ES_3 中	1732.00—1830.00	灰色泥岩
7	干 18	ES_4 上	1649.00—1655.00	灰色泥岩
8	阳 3	ES_3	1530.00—1630.00	油页岩
9	阳 4	ES_3	1962.00—1969.00	灰色泥岩
10	大 29	Ed_3	1900.00—1982.00	灰色泥岩
11	车 21	ES_1 上	1890.00—2000.00	灰色泥岩
12	大 26	ES_1 上	1850.00—1885.00	灰色泥岩

续表 7-11

样品号	井　号	层　位	井　段/m	岩　性
13	大 43	ES_4上	1933.00-1942.00	泥灰岩
14	垦 30	ES_1上	1900.00--1955.00	灰色泥岩
15	渔 23 开阔泻湖淤泥			
16	渔 41 半开口泻湖淤泥			
17	渔 33 封闭泻湖淤泥			
18	付疃河口泻湖淤泥			
19	付疃河口泻湖淤泥			
20	滦南 319 牛轭湖底淤泥			
21	微山湖底淤泥			
22	东海 8554 海底淤泥			
23	东海 8372、8232、8382 海底砂、粉砂			
24	渤 12#—1Q31.79—31.89 海底海相层淤泥质粘土			
25	渤 12#—1Q33.13—34.00 海底海相层淤泥质粘土			
26	渤 12#—1Q34.23—35.88 海底海相层淤泥质粘土			
27	渤中 12# 北 20 公里海底 23 米以内淤泥			
28	吉林双阳 40—130 厘米上黄色泥炭			
29	吉林双阳 130—165 厘米棕褐色泥炭			
30	吉林双阳 185 厘米以下黑腐值泥			

表 7-12　指标编号与名称对照表

指标号	指　标　名　称	指　标　号	指　标　名　称
1	"A" % 样品氯仿抽提物含量	15	C_{23}% 正烷烃
2	SA % 饱和烃	16	C_{24}%
3	AR % 芳烃	17	C_{25}%
4	NON-HC % 非烃	18	C_{26}%
5	AS% 沥青质	19	C_{27}%
6	C % 元素炭	20	C_{28}%
7	H % 氢	21	C_{29}%
8	O % 氧	22	C_{30}%
9	C_{17}% 正烷烃	23	PR/PH 姥值比
10	C_{18}%	24	$CPI=\frac{C_{23+25+27+29}}{C_{24+26+28+30}}$
11	C_{19}%	25	$WCI=C_{21+22}/C_{28+29}$
12	C_{20}%	26	$\sum C_{21}$前/$\sum C_{23}$后
13	C_{21}%	27	H/C 氢碳原子比
14	C_{22}%	28	O/C 氧碳原子比

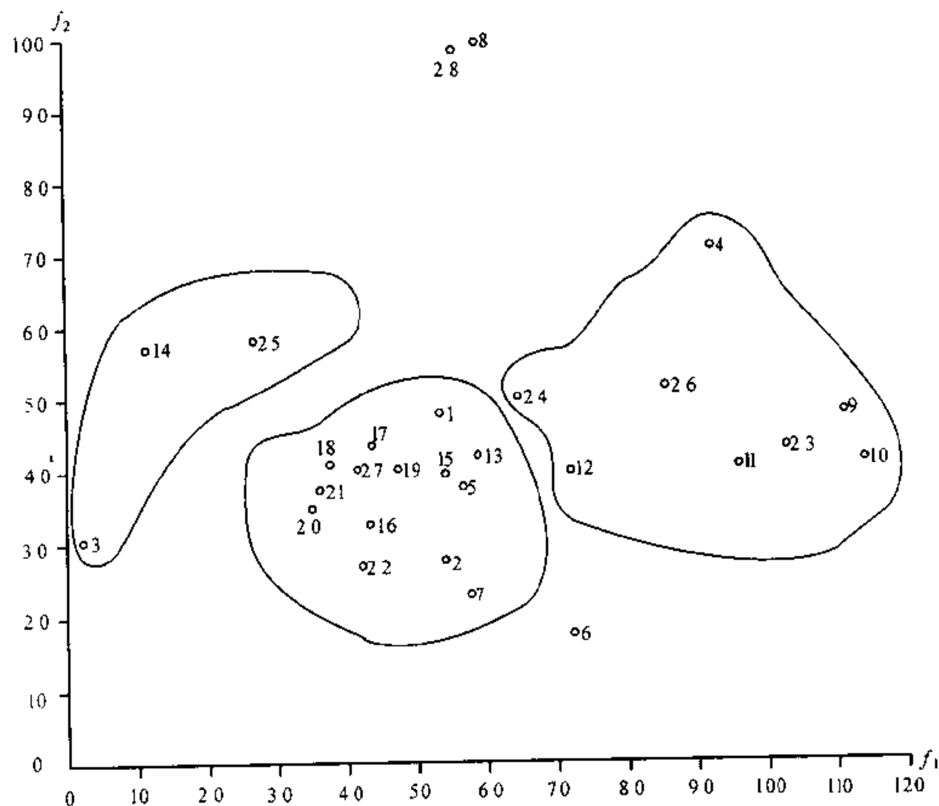


图 7-6 旋转后 f_1—f_2 因子载荷平面图

表 7-13 不同沉积相有机质地化指标比较表

样品类型	样品数	"A"/%	SA/%	AR/%	NON-HC/%	AS/%	O/%	$\sum C_{21}$前	C_{22}	$\sum C_{23}$后	PR/PH	CPI
吉林双阳炭土	3	1.525	3.1	0	59.5	36.96	11.30	25.6	6.8	37.4	0.63	3.2
微山湖滦南牛轭湖淡水淤泥	2	0.089	8.4	6.9	76.8	8.0	9.77	9.4	9.1	69.2	0.62	2.7
东营＜2000米生油岩	7	0.073	12.0	6.7	68.9	3.0	11.01	27.8	13.3	53.1	1.14	2.4
付疃河口堰塞湖淤泥	2	0.014	14.8	5.8	78.2	10.6	8.71	23.5	6.5	64.1	0.11	1.34
渔23等泻湖淤泥	3	0.034	15.6	24.7	56.1	4.1	10.17	13.5	32.1	48.8	0.14	1.61
渤海、东海、海底淤泥	6	0.031	12.7	26.4	59.5	1.45	10.29	6.3	27.0	57.9	0.21	1.5

对 30 个样品进行 Q 型因子分析。样品因子载荷平面图如图 7-7 所示。在图 7-7 上，样品分为三大组合区。

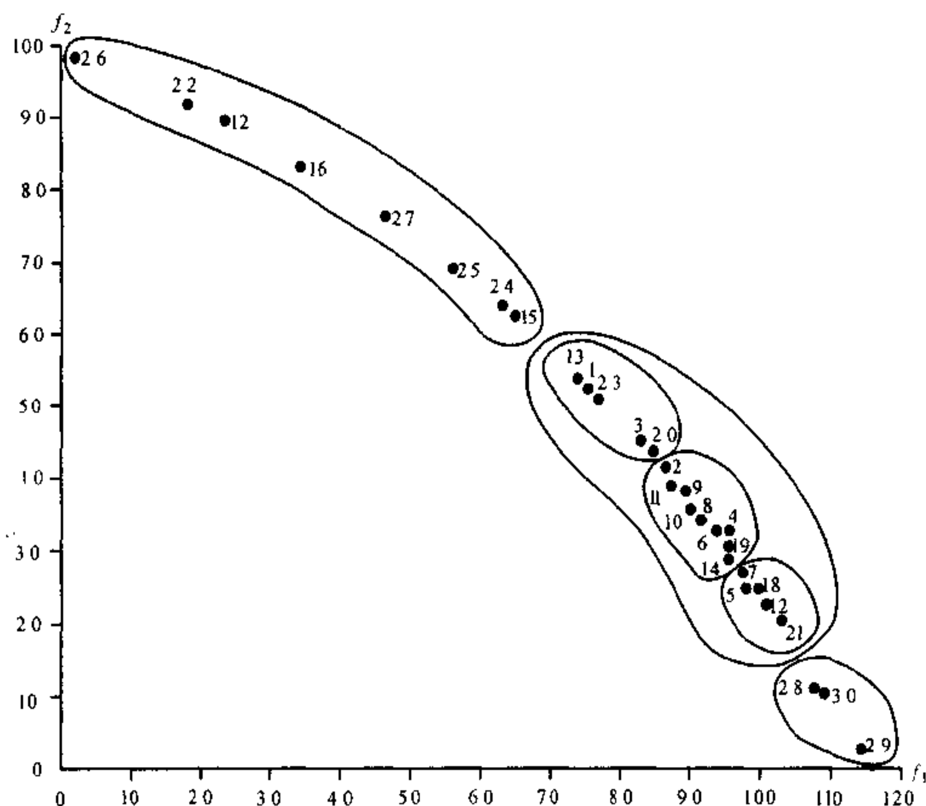


图 7-7 30块样品旋转后 f_1—f_2 因子载荷平面图

（据胜利油田地质科学研究院，1982，修改）

Ⅰ区位于图下部：为东北吉林双阳三块植物泥炭土(28、29、30 号)的组合，它与济阳坳陷诸样品、海相及内陆河湖样品显然分开，表明济阳坳陷生油岩与植物泥炭土的有机质是不相同的。

Ⅱ区为五块海底淤泥样品(22、24、25、26、27 号)和三块泻湖淤泥(15、16、17 号)的组合，这一组合表明海相和受海影响的泻湖淤泥的特征是相近的。这类有机质以水生生物为主，多为浮游动植物，富含蛋白质和脂类物质，少含木质素和纤维素，H/C 比值高，O/C 比值低，可能有较大的生油潜力。但这一组合中还有两个值得注意的问题，其一是三块泻湖淤泥样品，15 号为开阔泻湖淤泥，16 号为半开口泻湖淤泥，17 号为封闭泻湖淤泥，它们在散点图中的位置表明，同是泻湖沉积相有机质，而泻湖的类型不同时，有机质的性质尚有某种程度的差别。其二是 24、25、26 号样品均取自渤 12#—1 为第四系海相淤泥质粘土，然而 24、25 号样品相距较近，26 号样品相距较远，这表明它们的有机质性质有较大差别。从它们的原始数据上看，26 号样品的族组份上饱和烃大量减少，芳烃增多，非烃和沥青质下降，正烷烃 C_{22}上升，从元素组成上看，碳、氢、氧的含量总和 24 号为 95.20%，25 号为 91.60%，26 号为 90.82%，显示出规律性下降，这可能由于三块样品取样深度不同所致，亦可能与有机质来源有关。由于这些原因，使这一组合的各样品点，粗看起来类型一致，而实质上仍有较大的区别，这就是散点图中各样品点相距较远的原因。

Ⅲ区为济阳坳陷 14 块未成熟生油岩、付疃河口淤泥(28、29 号)、滦南牛轭湖(20 号)、微山

湖淤泥样品(21号)以及东海海底砂、粉砂(23号)样品的复杂组合。这一组合表明济阳坳陷生油岩有机质与植物泥炭土有机质有显著差异,而与淡水河湖相有机质相同,兼有海相和泻湖相有机质的特征。由于地质情况复杂,影响有机质分布的因素很多,根据样品点的分布,该组合可分为下述三个小组合进行研究:

① 1号(东营凹陷永31井沙三段上部)、13号(车镇凹陷大43井沙四段上部)与23号样品的组合。23号是东海近岸海底四块砂、粉砂样品的混合,岩性较粗,属近岸沉积物,有机质来自于陆源和海相沉积物,它与其他海底淤泥样品点相距较远,表明海相沉积物中不同亚相的有机质特征是不相同的。浅海亚相的泥质沉积物中以海洋浮游动植物为主,而滨海亚相的砂泥质混合沉积物中则具有海相和陆相有机质的共同特征。车镇凹陷大43井沙四段上部样品显示了与这种有机质相近的特征。根据车镇凹陷的沉积相研究,认为沙四段沉积时期,车镇凹陷一部份为咸化泻湖,大43井正处于这一湖区,虽然咸化湖蒸发量较大,海水进入受到局限,但紧临大海,地表升降运动频繁,海水仍可进入湖内,致使湖内生态变化受到海水影响,导致部份地区生油岩有机质与滨海沉积物相似。

② 东营凹陷的2、4、6号、惠民凹陷的8、9号、沾化凹陷的14号、车镇凹陷的10、11号样品、19号付疃河口淤泥、20号滦南牛轭湖淤泥样品的组合。这一组合内包括坳陷内所有四个次级凹陷的样品,并且这些样品又分布于主要生油层系的沙三段(2、4、6、8、9号)及沙一段(11、14号)和东营组(10号)各层段。这些样品点在图上密集分布,有的甚至互相叠合(为了标注样品号,图中把叠和样品分开了,但相应的位置没有改变),反映了这些有机质性质的高度相似性,这与济阳坳陷沉积发展史的统一性紧密相关。四个次级凹陷从沙三段到沙一段乃至东营组,基本上都为湖相沉积,具有相似的地质条件,生物群组合大体一致,提供的生油岩有机质基本相同。取自沾化凹陷垦30井沙一段上部的14号样品与付疃河口淤泥样品19号重合,并且与东营凹陷滨9井沙三段上部的5号样品及通23井沙三段中部的6号样品相近。同时取自车镇凹陷车21井沙一段上部的11号样品和惠民凹陷阳3井、阳4井沙三段样品8、9号重合为一点。这些样品的特征充分表明了这种沉积条件的统一性而导致生油岩有机质性质的相似性。然而,由于各凹陷有各自的物源区和沉积中心,不同时期沉积环境变化很大,虽同为湖相,各时期的沉积物大不一致,古生物及孢粉组合变化很大,氨基酸及有机碳的含量变化很不相同,因而不同地区、不同时代有机质的性质仍有很大差别,就是同一地区同一时代有机质性质亦有区别。例如,东营凹陷所取的1—7号样品,除7号(王18井)属于沙四段上部之外,其余均取自沙三段,除4号和6号重合之外,其余样品点并非聚集在一起,而且拉开较远,这就是统一性中的特殊性,它与这些样品取自湖泊中的不同地理位置有关。

③ 东营凹陷的5号(滨9井沙三段中部)、7号(王18井沙四段上部)和车镇凹陷的12号(大26号沙一段上部)样品与付疃河口18号,微山湖淡水湖相样品21号的组合。这一组合除了反映不同凹陷有机质性质的差异之外,还反映了同一凹陷沉积环境是渐变的。大26井沙一段上部的样品与微山湖前三角洲淤泥相近,表明大王庄地区沙一段沉积时期是淡水浅湖相沉积、与微山湖淤泥有机质性质相当。今日的微山湖水草丛生,芦苇繁茂,为陆源淡水湖相沉积环境,这可想象车镇凹陷沙一段时期的地理景观。将这一地区的沙四段与沙一段比较,它们的有机质体现出由近海相变成了湖相,而从散点图上看微山湖淤泥样品点近于东北吉林双阳泥炭土,可见随着陆源物质的加入及浅湖水生植物的繁茂而接近于泥炭土了。在这个组合中,取自东营凹陷滨9井沙三段下部的5号和王18井沙四段上部的7号样品与付疃河口淤泥18号样品重合。5号和7号样品取样位置相距不远,但分属于两段不同地层,这反映东营凹陷沙四段

末期与沙三段初期的沉积环境相同或渐变的。产自沙四上的艾氏鱼(*Knightia*)的生活环境是湿热的近岸海域或近海河湖，这种艾氏鱼在沙三下也有发现，说明沙四段上部和沙三段下部具有相似的生态环境，且不可排除这个时期有海水进入。

为了进一步研究各类沉积相样品与它们相应指标之间的关系，我们进行了对应分析，作出 f_1—f_2 平面散点图 7-8。图中表明样品的分类情况与 Q 型分析类似，但将各类型样品与其相应的特征指标反映到同一平面上来，它们间的相对关系可分为三区进行描述。其中Ⅱ区是济阳坳陷生油岩与淡水湖相和近海岸样品聚集区，特征指标为非烃、饱和烃、元素氧以及低碳部份和高碳部份正烷烃指标，显示出陆相生油岩有机质的特征。滨海湖盆中由于陆源植物和湖中低等生物有机质的充分供应，湖水营养价值很高，可滋生大量的生物，由于近海，海水不时侵入，同时得到海洋的补充，致使生物组合十分复杂。通过微体古生物化石的的研究，发现介形类、腹足类、藻类、轮藻、孢子花粉等种类繁多的生物群。东营凹陷始新世上部沙四段上部地层中局部地区见有海生的绿藻(*Chlorophyta*)、中国枝管藻(*Cladosiphonia sinensis*)和龙介虫(*Serpula*)与淡水的玻璃介(*Candona*)、金星介(*Cypris*)等相伴生形成混生生物群，可能由于这种复杂的生态体系，提供着济阳坳陷内生油母质。根据东营凹陷 13 块样品的干酪根分析，镜下鉴定统计了类脂组、壳质组、镜质组和惰性组的含量，计算出类型指数，确定干酪根也明显地属于混合型。

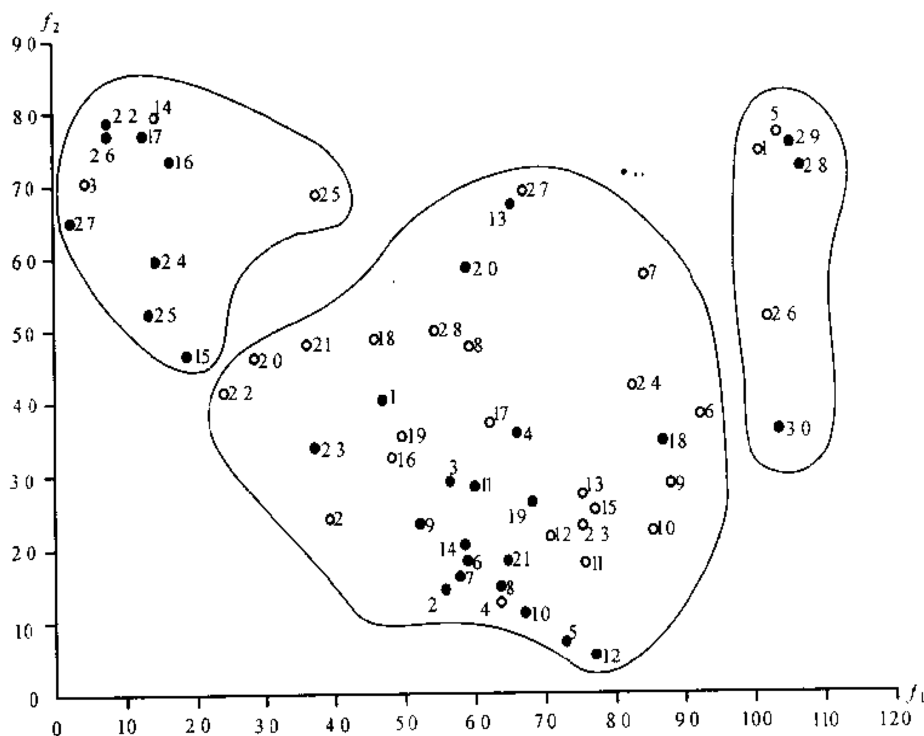


图 7-8　30 块样品 28 样指标对应分析图

● 样品　　○ 指标

因子分析和对应分析的结果都表明：济阳坳陷下第三系渐新统下部沉积有机质是混合型的，即以陆相断陷湖盆有机质为主，但亦有海生生物有机质的混入，表明了滨海块断湖盆沉积

面积成熟区的渐新世四子周围，说明此种沉积有机质控制油气分布。

习 题

1. 什么是因子分析？它在地质研究中的主要作用包括哪几个方面？
2. 因子分析目前在地质研究中的应用主要有哪些方面？
3. 什么是 R 型和 Q 型因子分析？
4. 写出 R 型和 Q 型因子分析模型，并说明 R 型因子模型中各个量的名称及统计意义？
5. 对 R 型和 Q 型因子分析如何求它们的主因子解？
6. 为什么要进行方差最大正交旋转？
7. 什么是因子得分？怎样计算因子得分？
8. 试述因子分析的计算步骤，如何应用各种计算结果？
9. 什么是对应分析？在对应分析图上可以获得哪些信息？
10. 试述对应分析的计算步骤。

第八章　地质有序数列分析

在地质研究中，有很多地质特征的观测值是一个有序的数据序列。例如沿垂直地质剖面或钻井岩心剖面上矿物的成分、气体的成分、颗粒的大小、地层的厚度、电性曲线等变化所形成的数列都是地质有序数列。由上述可知，地质有序数列是由描述某地质特征的数据构成的数列，将其简称为有序数列。一般将其记为

$$X = (x_1, x_2, \cdots, x_n) \tag{8-1}$$

其中 $x_i (i=1,2,\cdots,n)$ 为某地质特征的顺序观测值。

有序数列分析是研究有序数列间相互关系及性质的统计分析方法。这里介绍相关分析和趋势分析两种常用的有序数列分析方法 。

§1　相关分析

相关分析又分为互相关分析和自相关分析。前者研究二个或二个以上有序数列之间的相似性，而后者研究有序数列自身的性质。

一、互相关分析

地质上的很多对比问题都属于互相关分析的研究对象。例如，地层的对比就是互相关分析的典型例子。如果地层出露于地表，那么地层对比的问题不难解决。但是，地层的绝大部分，特别是有生油和储油意义的部分，都深埋于地下，在这种情况下，就要靠钻井取得的直接资料（岩心、岩屑）和间接资料（地球物理测井等）了解地层在地下的分布，由此产生了地质学中复杂的地层对比问题，即各井钻穿的地层间的相互关系是怎样的问题。解决这个问题的数学方法就是有序数列的互相关分析。其中的有序数列可以是钻井过程中获得的地层岩石样品在化验室测定的各种参数（变量）值，各种地球物理、地球化学测井数据等。下面以地层对比为例，介绍有序数列的互相关分析。

（一）简单地层剖面的对比

这里的简单地层剖面是指没有断层、地层尖灭、地层不整合面等复杂地质现象的地层剖面。

1．对比段的相似性

设 $X=(x_1, x_2, \cdots, x_G)$ 和 $Y=(y_1, y_2, \cdots y_N)(N \geqslant G)$ 是相邻两口井某测井参数构成的有序数列，那么地层对比就是找出 X 和 Y 中最相似的段，称为对比段。为此，可用互相关系数

$$R = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})^2 \sum_{k=1}^{n}(y_k - \bar{y})^2}} \tag{8 2}$$

作为对比段相关程度的指标。

这里的 n 是对比段内数据的个数，称其为对比段长度；$\bar{x}$、$\bar{y}$ 是有序数列 X、Y 在对比段内的数据平均值。

由式(8-2)可知：当两个有序数列在对比段内相同时，$R=1$；当 R 越接近于 $+1$ 时，两个有序数列在对比段上的关系就越密切，即两井中对应的地层段越相似；当 R 越趋向于 0 或 <0 时，对比段的差异就越明显。

由以上分析可知，根据各对比段上互相关系数 R 的值，便可确定相邻两井中最相似的地层段。

2. 对比过程

相邻两井有序数列互相关分析的过程如下：

(1) 把 x_i 与 y_j($i=1,2,\cdots,G$; $j=1,2,\cdots,N$; $N\geq G$)的头尾重迭 m 个数，如图 8-1a 所示，按式(8-2)计算对比段长度 $n=m$ 对比段的互相关系数 R_m。

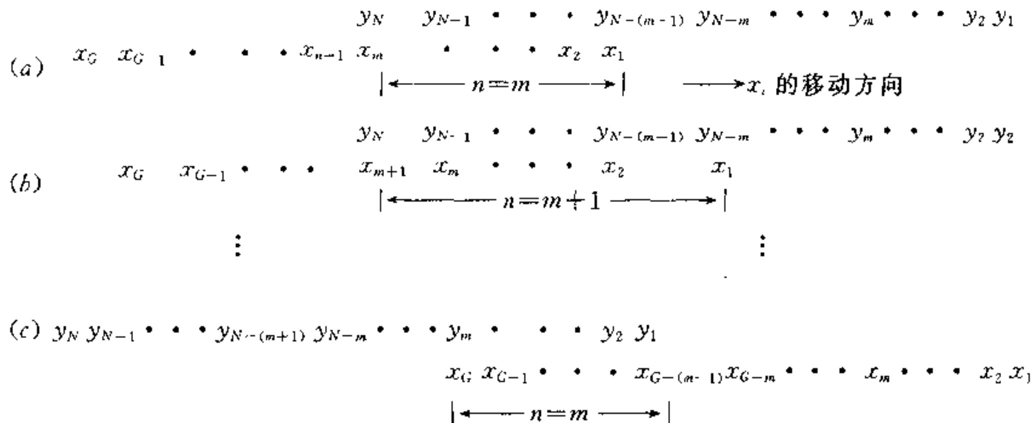


图 8-1 地层对比过程示意图

(2) 把有序数列 x_i 整体上向右移动一个数，如图 8-1b 所示，按式(8-2)计算对比段长度 $n=m+1$ 对比段的互相关系数 R_{m+1}。

⋮

重复上述做法，计算互相关系数，直到最终的对比段长度 $n=m$ 为止，如图 8-1c 所示。

完成整个对比过程后，可计算出 $P(P=N+G-2m+1)$ 个互相关系数。在简单地层剖面的条件下，显然，P 个互相关系数中的最大值所对应的对比段应是相邻两井地层剖面上最相似的地层段。计算机可以采用有序数列段或图形的方式输出最大相关系数所对应的对比段。

(二) 复杂地层剖面的对比

对具有断层、地层尖灭，不整合面等复杂地质现象的地层剖面进行对比就比较复杂。按简单地层剖面对比的方法，计算并绘制出相邻两井有序数列的互相关系数曲线，就会出现若干个峰值。每一个峰值都说明在它所对应的对比段上两个有序数列存在着关系密切的部分。在上述情况下，就要作仔细分析，研究这些相关段在地质上的意义。

现以具有断层的地层剖面对比为例，说明计算机对这种剖面对比的过程。相邻两井钻穿的地层剖面如图 8-2a 所示。对比这样的剖面时，就要找出两井剖面上的最大相关段、确定断点的位置和断层的落差等。

1. 寻找最大相关段

按照前面介绍的方法计算相邻两井两个有序数列在一系列对比段上的互相关系数 R，在用它绘制的曲线上，将不止一次地出现峰值。显然，当对比段在图 8-2b 及 C 的位置时，即 1 号井断层的上升盘地层与 2 号井相同地层对应及 1 号井断层下降盘地层与 2 号井相同地层对应

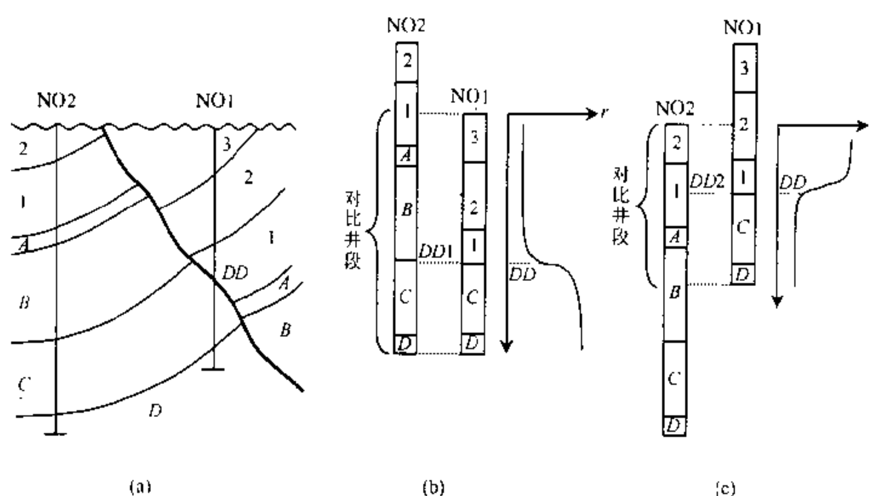


图 8-2 复杂地层剖面对比示意图

时，R 都会出现峰值，它们所对应的对比段应是最大相关段。如果把对比过程分为 NR 段，在每一段内找出 R 的峰值所对应的一个最大相关段，这样可以确定出整个对比过程中的 NR 个最大相关段，其中包括图 8-2b 及 C 位置时互相关系数的最大值所对应的最大相关段。

2. 断点和断距

在寻找出的 NR 个最大相关段内，它包含图 8-2b 及 C 位置时的两个最大相关段。现在我们分析这两个最大相关段的特点。在图 8-2 中，当两口井的对比段处在 b 的位置时，对比井段明显地分为两部分：下部地层相同，即 1 号井断层上升盘地层与 2 号井相对地层是同层；上部地层不同，即 1 号井断层下降盘地层与 2 号井相对地层不是相同的地层。根据这一特点，计算该段上的累加相关系数

$$r=\frac{\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k-\bar{x})^2\sum_{k=1}^{n}(y_k-\bar{y})^2}} \tag{8-3}$$

在形式上，式(8-3)与式(8-2)完全相同，但它的计算方法与式(8-2)不同。在计算 r 的过程中，最大相关段中的两个有序数列不作相对移动，仅是从最大相关段的上部开始，依次增大相关段的长度 n，计算相应的 r，直到使相关段长度 n 与最大相关段长度相等为止。

取 $n=n_0,n_0+1,\cdots\cdots$ 可计算出一系列的累加相关系数。对应于图 8-2b，我们分析一下 r 数列的特点：

从开始到断点 DD，因为最大相关段上部两口井对应部分是不同的地层，所以计算出的累加相关系数 r 将会很小。过断点后，再往下就是相同的地层，因此，从断点开始 r 值将会突然增大。按计算累加相关系数的顺序绘制 r 值变化的曲线时，在断点处将会出现一个明显的台阶，如图 8-2b 右所示。这个台阶指示的位置 DD 即为 1 号井内断点的位置。同理，在图 8-2c 位置时，在 r 值变化曲线上将会出现一个相反的台阶。它指示的位置也是 1 号井内的断点位置。计算机把 2 号井最大相关段上对应的断点深度 DD_1 及 DD_2 记录下来，两深度差即为断层的落差。

按上述分析，对 NR 个最大相关段计算累加相关系数 r，绘出 NR 条累加相关系数曲线，由这些曲线的特点可确定出相邻两井的对比层段、断层的断点和落差。

二、自相关分析

1. 自相关函数

自相关分析是研究有序数列自身周期性的一种数学方法。在式(8-2)中，如果有序数列 X 和 Y 是同一个有序数列，那么计算出的互相关系数就是有序数列自身的相关系数，因此，可用式(8-2)进行自相关分析。另外还可以定义自相关函数

$$r_t = \frac{\frac{1}{n-t}\sum_{k=1}^{n-t}(x_k - \bar{x})(x_{k+t} - \bar{x})}{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2} \tag{8-4}$$

$$(t = 1,2,\cdots,m(m \leqslant (n-1))$$

作为研究有序数列自身相互关系的数量指标，其中

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n}x_k$$

t ——有序数列的滞后项；

n ——有序数列的长度；

自相关函数具有以下性质：

(1) 当 $t=0$ 时，自相关函数有正的最大值，并在图形上呈对称图形；

(2) 当 $t\to\infty$ 时，$r_t\to 0$；

(3) 自相关函数 r_t 的波形与讯号本身波形无关，只与讯号本身的频率成分有关，即频率相同，波形不同的二种讯号可以有相同的自相关函数。

2. 相关图

相关图是以有序数列的自相关函数值 r_t 为纵坐标，以滞后项数 t 为横坐标绘制的曲线图，如图 8-3 所示。

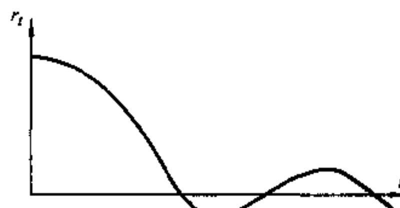


图 8-3　相关图示意图

利用相关图可以研究有序数列的周期性等特征。

§2　有序数列的趋势分析

在任何一个有序数列中，它的每一个观测值都由趋势变化、周期性变化和随机干扰三部分组成。其中对于解决地质问题有用的是趋势部分和周期性部分。经常采用滑动平均的方法，对有序数列进行处理。其目的在于：抑制原始数据中的噪声，把某一特征或变量在空间的变化分离为两部分，即趋势部分和剩余部分。对有序数列滑动平均，抑制随机干扰，提取有用讯号的过程称为有序数列的趋势分析。

滑动平均的具体做法是：在所要处理的图幅内，设置一个可以上下或左右滑动的“窗口”，把落在窗口内的全部数据进行平均，并把平均值置于窗口的中央，如图 8-4 所示。然后按一定的方向，把“窗口”或上或下，或左或右滑动一定的距离，于是，窗口便处于一个新的位置上，再把落在窗口内的全部数据进行平均，并把平均值置于新窗口的中央。以此类推，直到窗口滑遍

整个图幅为止。“窗口”内的全部数据可以按算术平均，也可以按不同的规则进行加权平均。

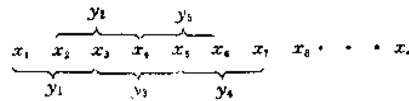


图 8-4　三点滑动平均过程示意图

较小“窗口”的滑动平均可以在一定程度上抑制原始数据中的噪声，较大“窗口”的滑动平均可以在一定程度上滤除剩余，突出趋势。滑动平均是一种简单的数字滤波方法，下面给出几个常用的滤波方程。

1. 3 项滤波方程

$$y_i = \frac{1}{3}(x_{i-1} + x_i + x_{i+1}) \ (i = 2,3,\cdots,n-1) \tag{8-5}$$

2. 5 项滤波方程

$$y_i = \frac{1}{35}(17x_i + 12(x_{i+1} + x_{i-1}) - 3(x_{i+2} + x_{i-2})) \ (i = 3,4,\cdots,n-2) \tag{8-6}$$

3. 7 项滤波方程

$$y_i = \frac{1}{21}(7x_i + 6(x_{i+1} + x_{i-1}) + 3(x_{i+2} + x_{i-2}) - 2(x_{i+3} + x_{i-3})) \quad (i = 4,5,\cdots,n-3) \tag{8-7}$$

4. 9 项滤波方程

$$y_i = 0.31x_i + 0.16(x_{i+1} + x_{i-1}) + 0.08(x_{i+2} + x_{i-2}) + 0.04(x_{i+3} + x_{i-3}) + 0.02(x_{i+4} + x_{i-4}) \quad (i = 5,6,\cdots,n-4) \tag{8-8}$$

5. 11 项滤波方程

$$y_i = \frac{1}{429}(89x_i + 84(x_{i+1} + x_{i-1}) + 69(x_{i+2} + x_{i-2}) + 44(x_{i+3} + x_{i-3}) + 9(x_{i+4} + x_{i-4}) - 36(x_{i+5} + x_{i-5})) \quad (i = 6,7,\cdots,n-6) \tag{8-9}$$

6. 15 项滤波方程

$$y_i = \frac{1}{320}(74x_i + 67(x_{i+1} + x_{i-1}) + 46(x_{i+2} + x_{i-2}) + 21(x_{i+3} + x_{i-3}) + 3(x_{i+4} + x_{i-4}) - 5(x_{i+5} + x_{i-5}) - 6(x_{i+6} + x_{i-6}) - 3(x_{i+7} + x_{i-7})) \quad (i = 8,9,\cdots,n-7) \tag{8-10}$$

7. 21 项滤波方程

$$y_i = \frac{1}{350}(60x_i + 57(x_{i+1} + x_{i-1}) + 47(x_{i+2} + x_{i-2}) + 33(x_{i+3} + x_{i-3}) + 18(x_{i+4} + x_{i-4}) + 6(x_{i+5} + x_{i-5}) - 2(x_{i+6} + x_{i-6}) - 5(x_{i+7} + x_{i-7}) - 5(x_{i+8} + x_{i-8}) - 3(x_{i+9} + x_{i-9}) - (x_{i+10} + x_{i-10})) \tag{8-11}$$

以上各式中的 x_i 是原始数据，y_i 是滤波后的新数据，n 为有序数列的数据总数，即原始有序数列的长度。

§3 有序数列分析 FORTRAN 源程序

一、相关分析程序

(一) 互相关分析程序

本程序模拟前节所述地层对比过程，结果以曲线图的形式给出并指示断点或其它复杂地质现象(当存在时)的位置。对程序中的主要参数、符号及程序的使用方法说明如下：

1. 主要参数及符号

(1) 参数

x0,y0——绘图坐标原点；

nin1——有序数列 1 的长度(采样点数)；

nin2——有序数列 2 的长度；

m1——有序数列 1 的起始深度，m；

m2——有序数列 2 的起始深度，m；

ndz——采样间隔值，m；

Lo——初始对比段长度；

ie——绘图控制参数，当 ie=0 时，全部结果绘制在一张图上，当 ie=1 时，每一个最大相关段绘制在一幅图上。

(2) 符号

xin1——存放有序数列 1 的数组名；

xin2——存放有序数列 2 的数组名；

fam1——有序数列 1 的数据文件名；

fam2——有序数列 2 的数据文件名；

mamin——求数列最大最小值子程序；

cor——求相关系数子程序；

lineo——绘制数列曲线子程序；

linr——绘制相关系数曲线子程序。

2. 程序使用说明

(1) 数据文件

在使用本程序时，先建立两个数据文件，其格式为：

$$(x_1\ x_2 \cdots x_{nin1})' \qquad (y_1\ y_2 \cdots y_{nin2})'$$

(2) 操作说明

在 DOS 操作系统下，键入互相关分析目标程序名 hxg 后的具体操作步骤如下：

① 输入绘图坐标原点(x0,y0)及采样间隔(ndz)(Input the origin of plotting coordinate (x0,y0) and ndz:)；

② 输入数列 1 和数列 2 的长度(Input the lendths of cruve 1 and curve 2 (nin1,nin2))；

③ 输入数列 1 和数列 2 的起始深度(Input the depths of starting point of curve 1 and curve 2 (m1,m2,(metre)：

④ 输入初始对比段长度(Input the original correlation interval (lo):)；

⑤ 输入绘图控制参数(Inpute the plotting control-Variable(ie):)；

⑥ 输入数列的数据文件名(Input fam1 fam2);操作完上述步骤后,程序运行结果,互相关分析结果存放在用户指定的图形文件中。

3. 源程序

(1) 源程序流程

源程序流程如图 8-5 所示。

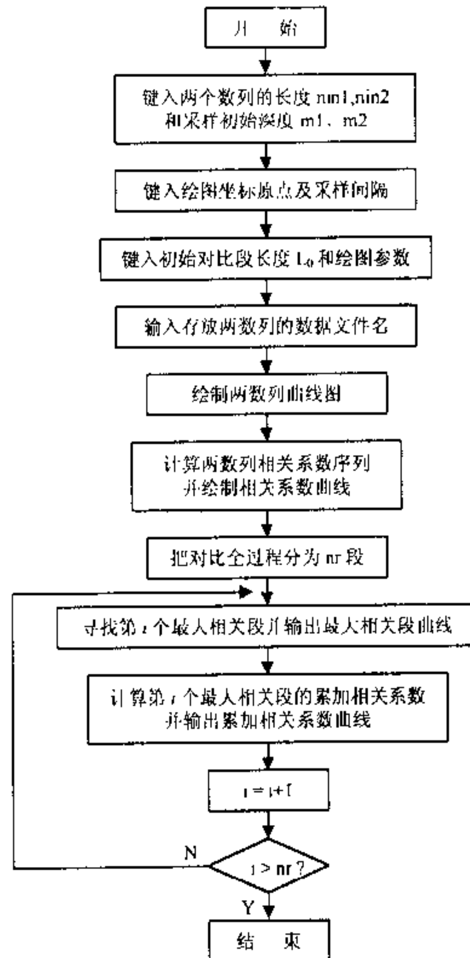


图 8-5 互相关分析流程图

(2) 源程序

```
$debug
        parameter (n1=1000,n2=1100,n3=2001)
        dimension ibc1(n3),ibc2(n3),rc(n3),xout(n3)
        dimension xinc1(n1),xinc2(n1),lenc1(n3)
        common/x1x2/xin1(n1),xin2(n2)
        common/lin/xs,xf,dz
        character fam1*10,fam2*10
        write(*,*) ' Input the origin of plotting '
```

```
      write( * , * ) ' coordinate(x0,y0) and dz: '
      read( * , * ) x0,y0,dz
      write( * , * ) ' Input the lendths of curve 1 '
      write( * , * ) ' and curve 2 (nin1=?,nin2=?): '
      read( * , * ) nin1,nin2
      write( * , * ) ' Input the depths of starting point '
      write( * , * ) ' of curve 1 and curve 2 (z1,z2,(metre)): '
      read( * , * ) z1,z2
      write( * , * ) ' Input the original correlation interval (lo): '
      read( * , * ) lo
      write( * , * ) ' Input the plotting control—variable(ie=0,1):'
      read( * , * ) ie
      write( * , * ) ' Input fam1,fam2 '
      read( * ,'(a)') fam1,fam2
      open(1,file=fam1)
      open(2,file=fam2)
c     open(3,file=' ')
      dh=0.5
      xs=0.
      xf=10.
      call in
      call fact(10.)
      call setsty('set10.sym')
      read(1, * ) (xin1(i),i=1,nin1)
      read(2, * ) (xin2(i),i=1,nin2)
      call mamin(xin1,nin1,ymin1,ymax1)
      call mamin(xin2,nin2,ymin2,ymax2)
      call bhdat(xin1,nin1,ymin1,ymax1)
      dy1=(ymax1-ymin1)/3
      call bhdat(xin2,nin2,ymin2,ymax2)
      dy=(ymax2-ymin1)/3
      call line(x0,y0,dh,dy,xin2,nin2,z2,ymin2)
      y10=y0+xf+8.
      call line(x0,y10,dh,dy1,xin1,nin1,z1,ymin1)
      not=nin1+nin2-2 * Lo+1
      ib1=1
      ie1=Lo
      ib2=nin2-Lo+1
      ie2=nin2
      len1=Lo
```

```
c       write(3,260)
        do 170 i=1,not
        call cor(len1,ib1,ib2,r)
        xout(i)=r
        br1=(ib1-1)*dz+z1
        er1=(ie1-1)*dz+z1
        br2=(ib2-1)*dz+z2
        er2=(ie2-1)*dz+z2
c       write(3,300) i,br1,er1,br2,er2,len1,r
        ibc1(i)=ib1
        ibc2(i)=ib2
        lenc1(i)=len1
        rc(i)=r
        ie1=ie1+1
        if(ie1-nin1) 120,120,100
100     ib1=ib1+1
        ie1=nin1
120     ib2=ib2-1
        if(ib2) 130,130,140
130     ib2=1
        ie2=ie2-1
140     len1=ie1-ib1+1
        len2=ie2-ib2+1
        if(len1-len2) 150,170,160
150     ib1=1
        ie2=ie2-1
        go to 140
160     ib1=ib1+1
        ie2=nin2
        go to 140
170     continue
        yr=y0+xf+xf+15.
        call linr(xout,x0,yr,dh,not,n3)
        nr=ifix(float(not)/Lo)
        xr=x0+0.2*not-10.
        do 230 j=1,nr
        k1=1+(j-1)*Lo
        k2=j*Lo
        rmax=1.e-02
        do 190 i=k1,k2
```

```
          if(rc(i).gt.rmax) go to 180
          go to 190
180       rmax=rc(i)
          ib1=ibc1(i)
          ib2=ibc2(i)
          len1=lenc1(i)
190       continue
          do 200 i=1,len1
          i1=i+ib1-1
          i2=i+ib2-1
          xinc1(i)=xin1(i1)
200       xinc2(i)=xin2(i2)
          if(ie.eq.1) then
          call in
          xr=x0
          end if
          br2=z2+ib2*dz-dz
          call line(xr,y0,dh,dy,xinc2,len1,br2,ymin2)
          yr=y0+xf+8.
          br1=z1+ib1*dz-dz
          call line(xr,yr,dh,dy1,xinc1,len1,br1,ymin1)
c         write(3,310)
c         write(3,330) rmax,br1,br2,len1
          k=0
          l=5
210       call cor(L,ib1,ib2,r)
          k=k+1
          xout(k)=r
c         write(3,320) k,l,r
          if(l.eq.len1) go to 220
          l=l+1
          go to 210
220       yr=yr+xf+7.
          call linr(xout,xr,yr,dh,len1,n3)
          if(ie.eq.0) xr=xr+0.2*len1+10.
230       continue
240       format(1x,' xmin1=',f8.4,' dx1=',f8.4)
250       format(1x,' xmin =',f8.4,' dx =',f8.4)
260       format(1x,' no ',' br1 ',' er1 ',
        # ' br2 ',' er2 ',' len1 ',' r ')
```

```
300       format(1x,i5,4f8.1,i8,f8.4)
310       format(6x,'rmax',6x,'ib1',7x,'ib2',7x,'len1')
320       format(5x,'k=',i5,' l=',i5,' r=',f10.4)
330       format(4x,f6.4,3x,f8.2,2x,f8.2,i8)
          stop
          end

          subroutine mamin(x,n,xmin,xmax)
          dimension x(n)
          xmin=x(1)
          xmax=x(1)
          do 100 i=1,n
          if(x(i).lt.xmin) xmin=x(i)
          if(x(i).gt.xmax) xmax=x(i)
100       continue
          return
          end

          subroutine bhdat(x,n,xmin,xmax)
          dimension x(n)
          common/lin/xs,xf,dz
          if(xs.ge.xf) then
          write(*,*) 'Data Error in Sub—bhdat ! '
          stop
          end if
          fc=(xf-xs)/(xmax-xmin)
          do 100 i=1,n
          x(i)=xs+(x(i)-xmin)*fc
100       continue
          end

          subroutine cor(n,ib1,ib2,r)
          common/x1x2/x1(1000),x2(1100)
          sx=0.
          sy=0.
          sxy=0.
          sxx=0.
          syy=0.
          do 100 i=1,n
          j1=ib1+i-1
```

```
          j2=ib2+i-1
          sx=sx+x1(j1)
          sy=sy+x2(j2)
          sxy=sxy+x1(j1) * x2(j2)
          sxx=sxx+x1(j1) * * 2
100       syy=syy+x2(j2) * * 2
          sq=sqrt((n * sxx-sx * sx) * (n * syy-sy * sy))
          r=(n * sxy-sx * sy)/sq
          end

          subroutine line(x0,y0,dh,dy,xi,n,z,ymin)
          dimension xi(n)
          common/lin/xs,xf,dz
          call movea(x0,y0)
          call linea(x0,y0+xf+2. )
          do 100 i=1,4
          y=ymin+(i-1) * dy
          y1=y0+(i-1) * (xf-xs)/3+xs
          call numb1(x0-0. 2,y1-1. 6,dh,90. ,y)
100       continue
          call movea(x0,y0+xi(1))
          x1=x0
          do 200 i=1,n
          call linea(x1,y0+xi(i))
200       x1=x1+0. 2
          call movea(x0,y0)
          call linea(x1,y0)
          x1=x0
          zz=y0-6 * dh-1.
          do 300 i=1,n
          x1=x1+0. 2
          if(mod(i,5). eq. 0) then
          zn=z+(i-1) * dz
          call numb1(x1,zz,dh,90. ,zn)
          end if
300       continue
          return
          end

          subroutine linr(xout,xr,yr,dh,not,n3)
```

```
      dimension xout(n3)
      x=xr
      call movea(x,yr-3.)
      call linea(x,yr+3.)
      call text(x-0.2,yr-40*dh,dh,90.,'-1')
      call text(x-0.2,yr-33*dh,dh,90.,'0 ')
      call text(x-0.2,yr-28*dh,dh,90.,'+1')
      call movea(x,yr)
      call linea(x+0.2*not,yr)
      call movea(x,yr+3*xout(1))
      do 100 i=1,not
      call linea(x,yr+3*xout(i))
100   x=x+0.2
      yyr=yr-4.
      x=xr
      call movea(x,yyr)
      do 200 i=1,not
      x=x+0.2
      if(mod(i,5).eq.0) then
      call numb2(x,yyr,dh,90.,i)
      end if
200   continue
      end
```

(二) 自相关分析程序

本程序对有序数列作自相关分析,结果以相关图的形式存放在用户指定的图形文件中。

1. 主要符号

x——存放数列滞后值的数组名;

y——存放数列值的数组名;

r——存放自相关函数值的数组名;

zxgt——绘制自相关图子程序;

sjbh——数据变换子程序;

bspli2——2 次 B 样条插值子程序;

sxs——画 x 轴子程序;

sys——画 y 轴子程序;

maxmin——求最大、最小值子程序。

2. 源程序

```
$debug
      common/xy/x(1000),y(1000),nx1,nx2,ny1,ny2
      dimension r(1000)
      common/max/xmax0,xmin0,ymax0,ymin0
```

```
        character fam1 * 10,fam2 * 10
        write( * , * ) ' Input your data file name:fam1,fam2 '
        read( * ,'(a)') fam1,fam2
        write( * , * ) 'Input xs,xf,ys,yf '
        read( * , * ) xs,xf,ys,yf
        write( * , * ) ' nx1,nx2,ny1,ny2 '
        read( * , * ) nx1,nx2,ny1,ny2
        open(1,file=fam1)
        open(2,file=fam2)
        do 100 i=1,10000
        read(1, * ,end=120,err=110) x(i)
100     read(2, * ,end=120,err=110) y(i)
110     write( * ,'(a,i5)') ' i=',i
        stop
120     n=i-1
        write( * ,'(a,i5)') ' End of data file: n=',n
        s=0.
        s1=0.
        do 130 i=1,n
        s=s+y(i)
130     s1=s1+y(i) * * 2
        s=s/n
        s1=s1/n-s * * 2
        do 150 i=1,n
        ss=0.
        j=i-1
        do 140 k=i,n
140     ss=ss+(y(k)-s) * (y(k-j)-s)
        r(i)=ss/s1/(n-j)
150     continue
        call in
        call fact(10. )
        call maxmin(x,n,xmax0,xmin0)
        call maxmin(r,n,ymax0,ymin0)
        xmax0=xmax0-1
        xmin0=xmin0-1
        call zxgt(r,n,xs,xf,ys,yf)
        stop
        end
```

```
      subroutine zxgt(r,n,xs,xf,ys,yf)
      dimension r(n)
      common/xy/x(1000),y(1000),nx1,nx2,ny1,ny2
      common/max/xmax0,xmin0,ymax0,ymin0
      common/og/dh,bh
      dh=0.2
      call sjbh(x,xs,xf,n)
      call sjbh(r,ys,yf,n)
      if(ymin0.lt.0.) y=ys-ymin0*bh
      if(ymin0.ge.0.) y=ys+ymin0*bh
      call bspli2(x,r,n,10)
      call sxs(xs,xf,y,nx1,nx2)
      call sys(ys,yf,xs,ny1,ny2)
      end

      subroutine sjbh(x,xs,xf,n)
      dimension x(n)
      common/og/dh,bh
      call maxmin(x,n,xmax,xmin)
      bh=(xf-xs)/(xmax-xmin)
      do 100 i=1,n
      x(i)=xs+(x(i)-xmin)*bh
100   continue
      end

      subroutine maxmin(x,n,xmax,xmin)
      dimension x(n)
      xmax=x(1)
      xmin=x(1)
      do 100 i=2,n
      if(x(i).gt.xmax) xmax=x(i)
      if(x(i).le.xmin) xmin=x(i)
100   continue
      end

      subroutine bspli2(x,y,n,nt)
      dimension x(n),y(n)
      dt=1.0/float(nt)
      call movea(x(1),y(1))
      x(1)=2*x(1)-x(2)
```

```
      y(1)=2*y(1)-y(2)
      x(n)=2*x(n)-x(n-1)
      y(n)=2*y(n)-y(n-1)
      do 20 i=1,n-2
      a0=0.5*(x(i)+x(i+1))
      a1=x(i+1)-x(i)
      a2=0.5*(x(i)-2*x(i+1)+x(i+2))
      b0=0.5*(y(i)+y(i+1))
      b1=y(i+1)-y(i)
      b2=0.5*(y(i)-2*y(i+1)+y(i+2))
      do 10 k=1,nt
      t=k*dt
      t2=t*t
      xx=a0+a1*t+a2*t2
      yy=b0+b1*t+b2*t2
      call linea(xx,yy)
10    continue
20    continue
      end

      subroutine sxs(xs,xf,y,n,ic)
      common/max/xmax0,xmin0,ymax0,ymin0
      common/og/dh,bh
      dx=(xf-xs)/n
      ddx=dx/ic
      dx1=(xmax0-xmin0)/n
      call movea(xs,y)
      call linea(xf+0.5,y)
      do 120 i=1,n+1
      x=xs+(i-1)*dx
      do 100 ii=2,ic
      xd=xs+(i-1)*dx+(ii-1)*ddx
      if(xd.gt.xf) go to 110
      call movea(xd,y)
      call linea(xd,y-0.1)
100   continue
110   call movea(x,y)
      call linea(x,y-0.2)
      xr=xmin0+(i-1)*dx1
      call numb1(x-3*dh,y-0.3-dh,dh,0.,xr)
```

```
120       continue
          end

          subroutine sys(ys,yf,x,n,ic)
          common/max/xmax0,xmin0,ymax0,ymin0
          common/og/dh,bh
          dy=(yf-ys)/n
          ddy=dy/ic
          dy1=(ymax0-ymin0)/n
          call movea(x,ys)
          call linea(x,yf+0.5)
          do 120 i=1,n+1
          y=ys+(i-1)*dy
          do 100 ii=2,ic
          yd=ys+(i-1)*dy+(ii-1)*ddy
          if(yd.gt.yf) go to 110
          call movea(x,yd)
          call linea(x-0.1,yd)
100       continue
110       call movea(x,y)
          call linea(x-0.2,y)
          yr=ymin0+(i-1)*dy1
          call numb1(x-7*dh,y,dh,0.,yr)
120       continue
          end
```

二、滑动平均程序

本程序按用户选定的滤波方程对一组原始数据进行数字滤波，抑制随机成分，保留趋势成分，并把滤波后的数据及其曲线图分别存放于用户指定的数据文件和图形文件中。

(一) 主要参数及符号。

xs,xf——数据 x 被变换到的端点值；

ys,sf——数据 y 被变换到的端点值；

x——存放原始数据 x 的数组名；

b——存放原始数 y 的数组名；

y——存放数据 y 滤波值的数组名；

f_{n1}. f_{n2}——存放原始数据 x,y 的数据文件名；

pointβ——β 项滤波方程子程序；

bspli2——2 次 B 样条插值子程序；

sxs——画 x 轴子程序；

sys——画 y 轴子程序。

(二)源程序

```
$debug
        dimension x(500),a(500),y(500)
        common/max/xmax,xmin,ymax,ymin
        character fn1*10,fn2*10
        write(*,*) 'Input xs,xf,ys,yf '
        read(*,*) xs,xf,ys,yf
        write(*,*) 'Input your data file name '
        read(*,'(a)') fn1,fn2
        open(1,file=fn1)
        open(2,file=fn2)
        open(3,file=' ')
        call in
        call fact(10.)
        call setsty("set10.sym")
        do 100 i=1,10000
        read(1,*,end=120,err=110) x(i)
100     read(2,*,end=120,err=110) a(i)
110     write(*,*) 'i=',i
        stop
120     n=i-1
        write(*,'(a,i4)') 'End of file n=',n
        write(*,*) 'Input m '
        read(*,*) m
        call maxmin(x,n,xmax,xmin)
        call maxmin(a,n,ymax,ymin)
        call sjbh(x,xs,xf,n)
        call sjbh(a,ys,yf,n)
        if(m.eq.3) call point3 (a,y,n)
        if(m.eq.5) call point5 (a,y,n)
        if(m.eq.7) call point7 (a,y,n)
        if(m.eq.9) call point9 (a,y,n)
        if(m.eq.11) call point11(a,y,n)
        if(m.eq.15) call point15(a,y,n)
        if(m.eq.21) call point21(a,y,n)
        i1=(m-1)/2+1
        i2=n-(m-1)/2
        write(3,'(1x,f8.2)') (y(i),i=i1,i2)
        if(m.eq.0.) call bspli2(x,a,i1,i2,10)
```

```
      if(m.eq.3.or.m.eq.5.or.m.eq.7.or.m.eq.9.or.m.
     # eq.11.or.m.eq.15.or.m.eq.21) call bspli2(x,y,i1,i2,10)
      call sxs(xs,xf,ys,0.2,10,10)
      call sys(ys,yf,xs,0.2,5,5)
      stop
      end

      subroutine point3(a,b,n)
      real a(n),b(n)
      do 100 i=2,n-1
      b(i)=(a(i-1)+a(i)+a(i+1))/3.0
100   continue
      b(1)=0.5*(a(1)+a(2))
      b(n)=0.5*(a(n-1)+a(n))
      end

      subroutine point5(a,b,n)
      dimension a(n),b(n)
      do 100 i=3,n-2
      b(i)=(17*a(i)+12*(a(I+1)+a(i-1)) -3*(a(i+2)+a(i-2)))/35
100   continue
      end

      subroutine point7(a,b,n)
      dimension a(n),b(n)
      do 100 i=4,n-3
      b(i)=(7*a(i)+6*(a(i+1)+a(i-1))
     # +3*(a(i+2)+a(i-2))-2*(a(i+3)+a(i-3)))/21
100   continue
      end

      subroutine point9(a,b,n)
      dimension a(n),b(n)
      do 100 i=5,n-4
      b(i)=0.31*a(i)+0.16*(a(i+1)+a(i-1))+0.08*(a(i+2)
     # +a(i-2))+0.04*(a(i+3)+a(i-3))+0.02*(a(i+4)+a(i-4))
100   continue
      end

      subroutine point11(a,b,n)
```

```
      dimension a(n),b(n)
      do 100 i=6,n-5
      b(i)=(89*a(i)+84*(a(i+1)+a(i-1))+69*(a(i+2)
     # +a(i-2))+44*(a(i+3)+a(i-3))-9*(a(i+4)
     # +a(i-4))-36*(a(i+5)+a(i-5)))/429
100   continue
      end

      subroutine point15(a,b,n)
      dimension a(n),b(n)
      do 100 i=8,n-7
      b(i)=(74*a(i)+67*(a(i+1)+a(i-1))+46*(a(i-2)+a(i-2))
     # +21*(a(i+3)+a(i-3))+3*(a(i+4)+a(i-4))-5*(a(i+5)
     # +a(i-5))-6*(a(i+6)+a(i-6))-3*(a(i+7)+a(i-7)))/320
100   continue
      end

      subroutine point21(a,b,n)
      dimension a(n),b(n)
      do 100 i=11,n-10
      b(i)=(60*a(i)+57*(a(i+1)+a(i-1))+47*(a(i+2)+a(i-2))
     # +33*(a(i+3)+a(i-3))+18*(a(i+4)+a(i-4))+6*(a(i+5)
     # +a(i-5))-2*(a(i+6)+a(i-6))-5*(a(i+7)+a(i-7))
     # -5*(a(i+8)+a(i-8))-3*(a(i+9)+a(i-9))
     # -(a(i+10)+a(i-10)))/350
100   continue
      end
```

§4 应用算例

【例 1】 这是两口相邻井的部分对比电测曲线的对比实例。1 号井曲线包含 240 个采样值,2 号井曲线包含 329 个采样值。取样间隔为 1 米,其中 1 号井通过一条断层。

两条曲线上的采样值分别见表 8-1 和表 8-2

表 8-1 1 号井电测曲线采样值(每行从左至右)

2.1	2.3	2.4	2.2	2.2	2.2	2.4	2.2	4.2	9.7
1.9	2.9	2.0	2.2	2.8	1.9	2.1	2.0	1.8	1.7
1.8	1.6	1.6	1.5	1.3	1.4	1.6	.9	1.4	1.9
1.6	1.6	1.5	1.6	1.4	1.5	1.6	1.7	1.6	1.5
1.5	1.5	1.4	1.5	1.7	1.8	2.2	2.8	1.6	2.0
1.9.	2.0	1.8	2.4	2.0	2.1	1.8	1.6	1.6	1.7
1.3	1.5	1.4	1.5	2.0	2.1	2.6	3.0	2.1	2.4

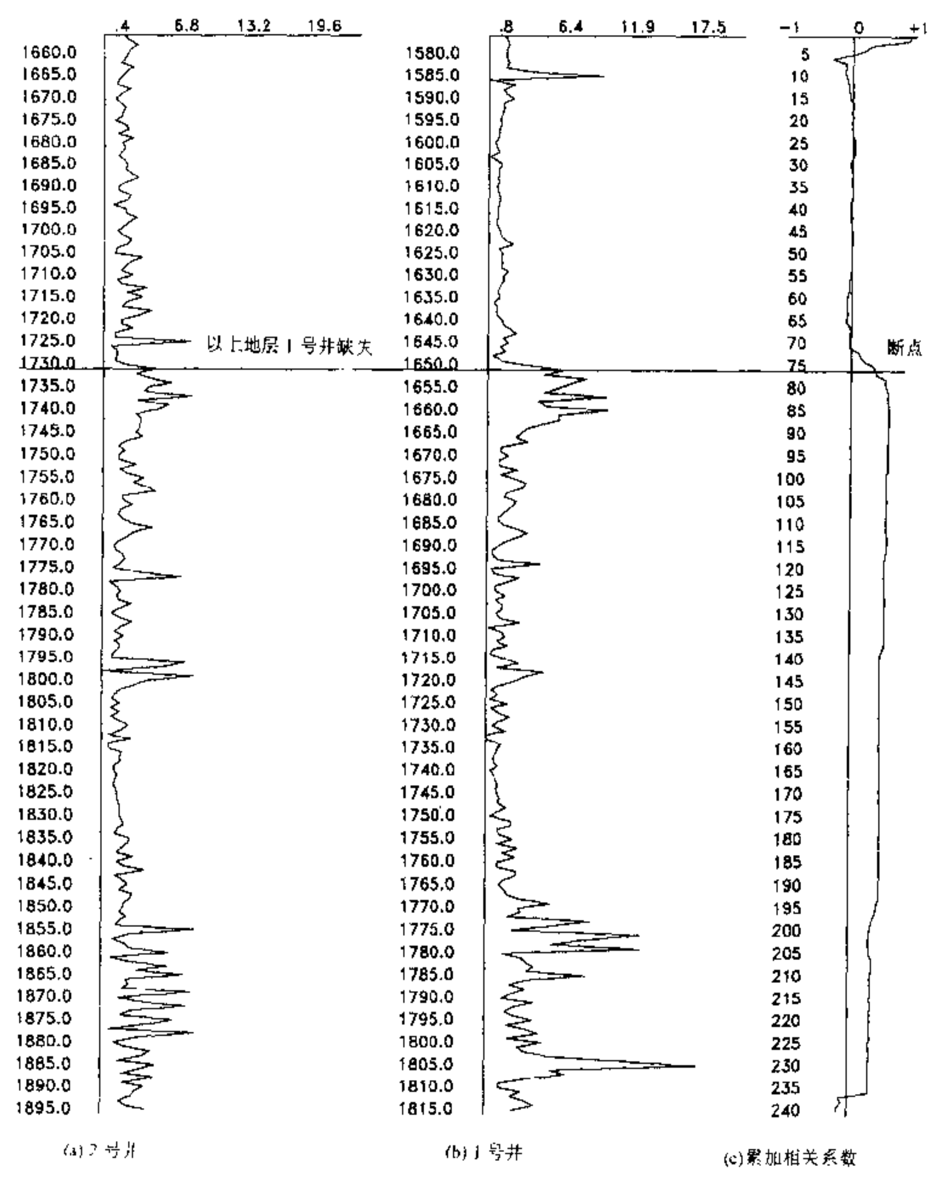
续表 8-1

1.9	2.2	1.2	2.0	4.0	6.8	5.2	8.6	7.4	6.4
5.3	10.1	4.8	5.4	10.2	6.4	6.5	5.2	4.0	3.5
3.1	4.0	2.0	1.9	2.5	1.8	1.7	3.2	2.0	3.0
3.9	3.6	2.4	2.2	3.1	2.9	2.4	2.0	1.7	2.1
3.1	4.0	3.2	2.2	1.7	1.2	1.3	1.7	5.0	1.2
1.4	3.4	2.4	1.4	1.5	2.3	1.5	2.1	2.1	1.6
2.0	3.5	1.0	2.1	2.3	1.6	2.5	1.3	1.2	2.4
3.4	1.7	5.3	3.8	3.0	1.8	1.2	1.8	1.1	2.6
1.2	1.8	2.7	1.9	1.3	2.4	.8	.9	2.1	1.8
1.7	1.4	1.4	1.8	1.2	1.4	1.6	1.8	1.7	1.8
2.0	1.8	2.5	1.8	1.2	2.8	2.9	1.8	1.9	3.0
1.8	3.4	2.2	3.2	1.9	1.7	3.2	3.2	2.2	2.0
2.4	2.6	3.6	5.9	3.5	3.6	2.6	9.0	6.0	3.0
13.0	9.0	6.0	13.0	2.2	3.7	4.0	4.6	4.1	8.8
4.2	2.8	4.0	3.0	2.2	4.6	3.0	2.4	3.2	4.6
2.6	4.0	5.0	2.6	5.3	3.4	4.0	6.0	12.0	17.5
6.0	7.2	4.2	2.0	2.0	3.7	2.9	4.0	4.6	3.0

表 8-2　2号井电测曲线采样值(每行从左至右)

5.4	3.5	2.8	3.8	2.8	3.1	3.1	2.8	3.0	8.0
6.4	1.6	4.7	2.1	2.6	2.0	2.1	2.1	2.0	2.2
2.1	2.0	2.1	2.0	2.4	2.0	1.7	2.4	2.0	1.9
2.0	2.0	2.0	2.1	1.6	1.9	2.0	1.9	1.7	2.1
2.0	2.1	2.9	3.9	4.2	2.4	2.8	2.6	2.2	1.8
2.0	1.8	2.0	1.8	1.8	2.2	2.1	1.6	2.0	1.2
1.4	1.5	4.0	3.7	2.3	3.4	6.8	5.2	10.0	20.0
12.4	18.8	2.2	3.3	3.6	2.6	3.6	2.0	1.8	2.4
1.8	2.0	1.8	2.0	1.8	2.2	2.6	3.4	2.7	2.5
2.2	1.9	3.0	2.5	2.1	1.7	1.5	2.8	2.0	1.4
1.8	2.3	2.1	2.0	1.4	2.3	2.6	1.6	3.0	2.0
2.2	2.0	1.7	2.2	2.5	2.3	2.8	3.5	2.4	1.6
1.7	2.8	2.6	1.2	2.3	2.4	3.3	2.6	2.1	1.6
2.0	2.8	2.6	1.5	1.4	3.9	2.8	2.3	1.9	3.0
1.6	1.9	4.4	2.5	4.0	2.6	2.0	4.8	3.6	2.0
2.0	3.0	1.4	1.5	8.4	1.1	1.6	1.6	1.4	2.4
5.2	3.3	4.6	6.6	4.6	4.2	8.4	3.5	6.4	5.4
3.3	3.8	3.8	3.7	3.3	4.0	2.4	1.8	2.3	1.8
1.8	3.4	1.9	2.6	4.2	3.1	4.4	5.2	2.1	2.2
3.0	2.0	1.8	2.2	3.0	5.0	3.2	2.6	1.8	1.4
1.6	2.2	2.4	2.0	1.3	4.0	7.6	1.0	1.6	2.0
1.7	1.8	3.0	2.0	1.2	1.9	3.1	2.0	1.5	2.3
1.5	2.0	1.8	1.4	1.3	8.0	6.0	.4	8.8	4.6
3.4	2.0	1.5	1.3	2.0	1.2	2.0	1.2	2.0	2.8
2.1	1.0	3.0	1.0	1.0	2.2	2.0	2.1	1.5	1.5
1.8	1.6	1.4	1.7	1.8	1.8	1.9	2.0	2.0	2.0
2.2	2.4	1.8	1.6	3.0	2.8	2.2	1.8	3.2	1.8
2.4	4.4	1.6	2.3	2.8	2.0	3.2	3.0	2.4	2.4
2.0	2.6	2.0	1.7	9.0	3.0	1.4	2.6	3.0	6.6
1.2	3.0	6.4	3.7	8.0	2.6	2.9	2.4	8.6	2.0
4.0	8.2	3.0	2.2	7.0	5.0	1.1	9.0	3.6	1.6
2.6	5.0	4.2	2.0	5.4	4.0	2.2	5.3	2.1	1.4
3.6	3.2	2.0	2.6	4.5	1.6	1.6	1.6	2.9	

对比结果由两条曲线表示的地层的对应位置如图 8-6 所示。从累加相关系数曲线上看出，在序号为 70 处出现台阶。但是，因为我们是从 5 个对应点开始计算的累加相关系数，因此，断点的位置应在序号为 75 处，即 1 号井 1650 米深处。相当于 2 号井深度为 1730 米以上的地层在 1 号井缺失。



(a) 2 号井　　(b) 1 号井　　(c) 累加相关系数

图 8-6　打印机输出的对比结果曲线

【例 2】 美国绿河(Green River)油页岩季候泥分层厚度数据的自相关分析。

分层厚度数据见表 8-3

表 8-3　绿河油页岩季候泥分层厚度数据

数据总数/101			单位/mm				采样时间间隔/a		
6.0	7.2	7.1	7.1	7.2	7.4	8.0	8.6	10.0	12.4
12.0	11.0	9.6	8.7	7.6	7.2	7.2	7.8	8.1	7.8
7.1	7.2	7.1	7.0	7.0	7.7	8.6	9.0	12.0	13.7
14.0	13.6	12.1	12.9	12.8	11.1	9.0	7.5	7.5	8.4
8.4	7.9	7.0	6.7	6.8	7.5	7.3	7.2	8.1	9.8
11.0	10.8	9.5	8.1	7.2	7.1	6.8	7.0	7.1	5.6
3.8	3.4	4.2	4.8	4.5	3.6	3.0	2.8	4.1	6.8
8.1	7.8	6.4	4.6	3.7	4.0	4.2	4.5	5.9	7.3
7.3	6.7	6.0	5.8	8.7	6.5	8.2	10.2	12.3	13.2
13.2	12.4	9.7	9.2	9.3	8.3	6.0	5.7	6.1	6.3
6.3									

由自相关图可以看出，季候泥分层厚度数据序列含有 20 年的周期，这接近太阳黑子活动的 22 年周期，说明季候泥分层厚度的变化与此有关。

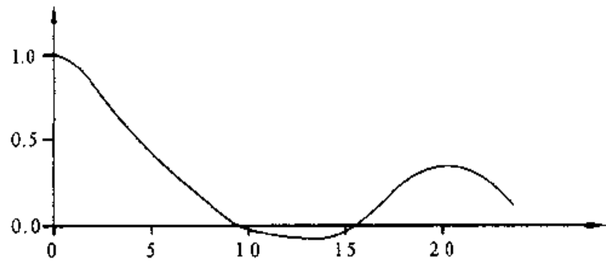


图 8-7　油页岩季候泥分层厚度数据序列自相关图

【例 3】 利用滑动平均抑制有序数列中的随机成分。

利用 5 项和 7 项滤波方程对表 8-3 中数据进行数字滤波，结果见图 8-8。由图 8-8 可见，随着滤波方程项数的增加，曲线变的更加光滑。

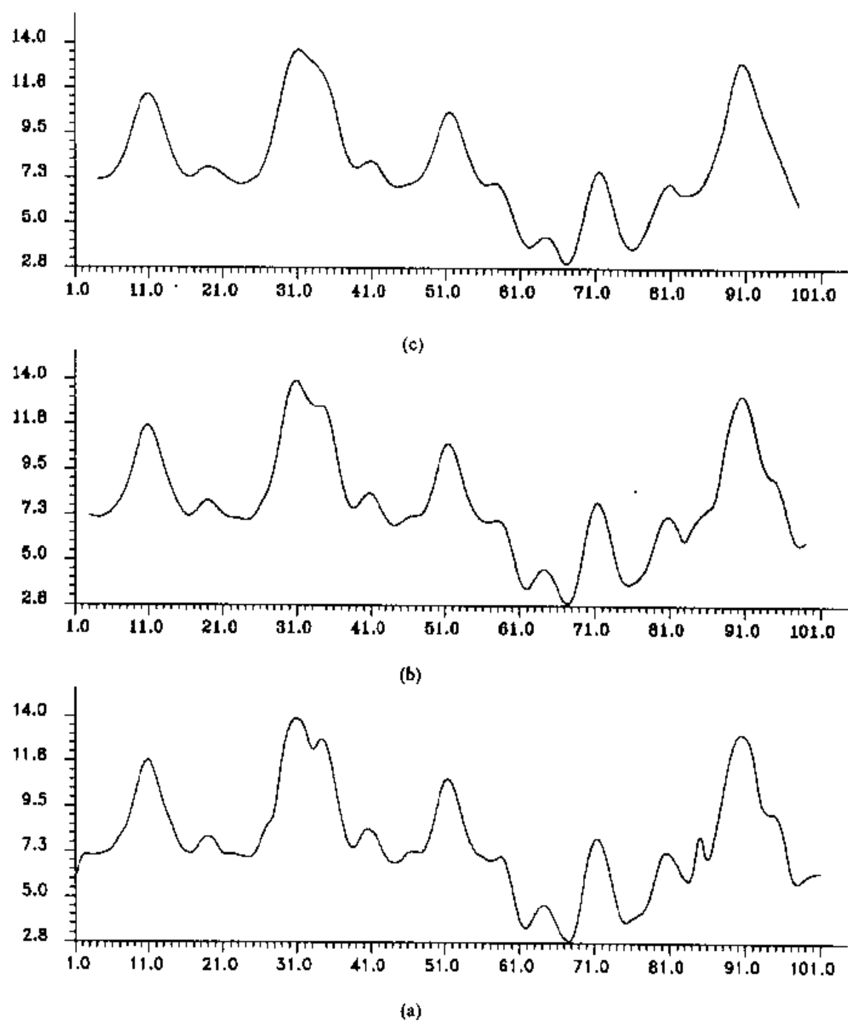


图 8-8　滤波后的曲线图

(a) 原始数据曲线；(b) 5 项方程滤波后曲线；(c) 7 项方程滤波后曲线

习　题

1. 什么是有序数列分析?
2. 应用互相关和自相关分析可以研究地质学中哪一类的问题?
3. 试述相关分析的过程。
4. 何谓有序数列的趋势分析?
5. 试述滑动平均法抑制随机干扰的过程。
6. 试举例说明有序数列分析在石油及天然气地质中的应用。

第九章　马尔可夫概型分析

任何一个地质过程都可认为是某些确定性地质因素和随机性地质因素相互作用的产物。也就是说，地质过程都应是确定型和随机型地质过程在时间上和空间上叠加的结果，因此，在地质研究工作中，应重视上述两类地质过程的作用。但是，由于产生随机地质过程的机理十分复杂，甚至连随机性地质因素也不容易或不可能观察，这就给用数学模型（一个或几个）来描述随机型地质过程带来了极大的、甚至是在目前条件下不可能克服的困难。这样以来，人们自然就会侧重于确定型地质过程的研究，而使随机型地质过程的研究成为地质工作中的一个薄弱环节，到目前为止，研究随机型地质过程所涉及的数学方法也仅限于马尔可夫概型分析。

1949年，维斯捷列乌斯在研究复式沉积层形成问题时，首先应用了马尔可夫链。本世纪60年代，有关马尔可夫过程在地质学中应用的论文大量涌现。70年代以后虽然论文数量有所下降，但人们对其理解程度却在加深，解决实际问题的效果不断提高，应用领域也有所扩大。

1984年，在莫斯科举行的第27届国际地质会议上，维斯捷列乌斯、阿格特伯格（F.P. Agterbers）等数学地质学家发表了莫斯科宣言，其基本思想是：随机型模型应作为数学地质模型的基础，并且肯定了马尔可夫过程（特别是马尔可夫链）在地质学中应占有特殊的地位。

目前，常用马尔可夫概型分析研究沉积旋回，进行地层对比，查明火山岩系的喷出顺序和侵入杂岩体中各个侵入体形成的先后顺序，划分矿床的成矿期和成矿阶段、揭示各个成矿阶段的空间分布等。

§1　马尔可夫概型

随机过程是概率论的基本概念之一，它是依赖于参数 t 的一族随机变量，记为

$$\{x(t), t \in T\} \tag{9-1}$$

这里的参数 t 一般是时间，T 是它的变化范围，随机变量 $x(t_i)$ 也称作随机过程 $x(t)$ 在 $t=t_i \in T$ 时的状态。

当随机过程在时刻 t_1 所处的状态 $x(t_1)$ 为已知的条件下，若随机过程在时刻 $t(t>t_1)$ 所处的状态 $x(t)$ 与随机过程在 t_1 时刻之前发生的状态无关，那么这样的随机过程就称为马尔可夫过程。用分布函数来描述就是：如果对于参数 t 的任意 $n(n\geqslant 3)$ 个数值 $t_1<t_2<\cdots<t_n$，在条件 $x(t_i)=x_i, i=1,2,\cdots,n-1$ 下随机变量 $x(t_n)$ 的分布函数恰好等于在条件 $x(t_{n-1})=x_{n-1}$ 下 $x(t_n)$ 的分布函数，即

$$\begin{aligned} & F(x_n;t_n/x_{n-1},x_{n-2},\cdots,x_1;t_{n-1},t_{n-1},\cdots,t_1) \\ & = F(x_n;t_n/x_{n-1};t_{n-1}) \qquad (n=3,4,\cdots) \end{aligned} \tag{9-2}$$

则称 $x(t)$ 为马尔可夫过程。条件分布函数

$$F(x_n;t_n/x_{n-1};t_{n-1}) = P\{x(t_n) \leqslant x_n/x(t_{n-1}) = x_{n-1}\} \quad (t_n > t_{n-1}) \tag{9-3}$$

是马尔可夫过程的概率模型，称式(9-3)为马尔可夫过程的转移概率。

分布函数式(9-2)给出了马尔可夫过程的概率特性，称这一特性为无后效性，因此，马尔可夫过程又称为"无后效随机过程"。所谓无后效性就是在已知随机过程现在状态的情况下，以后

它所处的状态与以前它所处的状态无关。可以将其理解为，这个过程的历史对未来的全部影响集中在最后时刻的状态中，即认为过程的任何观测结果只与紧前面的观测结果有关。

状态和参数都是离散的马尔可夫过程称作马尔可夫链，即过程为

$$\{x_t, t = 0,1,2,\cdots\} \tag{9-4}$$

这里状态的数目可以是有限的或可列无穷的。

马尔可夫链适用于时间离散、状态离散的时间序列。但是，在研究地质过程时，有时能直接确定过程在时间上的先后顺序，有时则只能间接地以空间上的上下、前后、左右关系来代替；也就是说，在地质过程研究中，有时可以找到确定的时间序列，有时只能间接地用距离来代替时间参数。但是，只要空间序列有类似于马尔可夫性质的关系存在，则仍然可以应用马尔可夫链对这种序列进行研究。在此，我们将既适用于时间序列又适用于空间序列的马尔可夫概率模型统称为“马尔可夫概型”。

若马尔可夫过程的转移概率随着时间的推移而发生变化，则称其为非齐次或非平稳马尔可夫过程，转移概率不随时间而变的马尔可夫过程就称为齐次或平稳马尔可夫过程。目前在地质研究中，主要是应用平稳马尔可夫过程。

§2　马尔可夫链的转移概率

一、一阶转移概率

设马尔可夫链中可列个发生状态转移的时刻为 $t_1, t_2, \cdots t_n, \cdots$，在已知时刻 $t=t_n$ 时随机过程 x_t 所处状态为 i 的条件下，把经过一步转移，即在时刻 $t=t_{n+1}(t_{n+1}>t_n)$ 转移到状态 j 上的概率记为 p_{ij}，相应于式(9-2)则有

$$p_{ij} = p\{x_{t_{n+1}} = j / x_{t_n} = i\} \tag{9-5}$$

这个概率称为马尔可夫链的一阶转移概率。为了明确起见，把从状态 i 到状态 j 的一阶转移概率 p_{ij} 记为 $p_{ij}^{(1)}$，…，而 $p_{ij}^{(k)}$ 则是从状态 i 出发经过 k 步转移到状态 j 上的转移概率 。如果 $p_{ij}^{(k)}$ 只与状态 i、j 及转移步数 k 有关，而与具体那个时刻无关，这时称马尔可夫链为平稳的或齐次的。严格地讲，这种仅与最后状态有直接关系的马尔可夫链称为一重马尔可夫链。如果一个状态转移的条件概率不仅与前面一个，而且与前面两个，甚至是几个状态有关时，则称这样的马尔可夫链为二重或几重马尔可夫链。下面主要是讨论一重马尔可夫链。

对于马尔可夫链来说，转移概率完全描述了它的概率统计特征，因此，如何确定转移概率则成为研究马尔可夫链的一个重要问题。转移概率在理论上是条件概率，而实际应用时则是以转移频率 $n_{ij}/n_{i\cdot}$ 作为条件概率的估计值，即

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{i\cdot}}$$

其中 $n_{i\cdot}$ 是状态 i 出现的次数，n_{ij} 是从状态 i 一步转移到状态 j 的次数。

例如，若在某个地层剖面中岩性这个随机变量只能取砂岩(E_1)，粉砂岩(E_2)和泥岩(E_3)三种状态，对剖面中的岩性变化从底到顶观测记录如下：

$$E_1\ E_1\ E_2\ E_1\ E_3\ E_2\ E_2\ E_1\ E_1\ E_2\ E_3\ E_3\ E_1\ E_2\ E_3\ E_1$$

在某次观测状态为 E_i 的条件下，下次观测为 E_j 的条件概率记为 $p_{ij}^{(1)}$，其中 i，j 分别表示起始和终止状态。在上面列出的地层剖面中，砂岩(E_1)出现了 7 次，最后一次出现砂岩的后面已无资料，所以以砂岩为起始状态来统计下次出现什么状态只能统计 6 次。经统计得出：

$$P^{(1)}=\begin{pmatrix}p_{11}^{(1)} & p_{12}^{(1)} & p_{13}^{(1)}\\ p_{21}^{(1)} & p_{22}^{(1)} & p_{23}^{(1)}\\ p_{31}^{(1)} & p_{32}^{(1)} & p_{33}^{(1)}\end{pmatrix}=\begin{pmatrix}\frac{2}{6} & \frac{3}{6} & \frac{1}{6}\\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5}\\ \frac{2}{4} & \frac{1}{4} & \frac{1}{4}\end{pmatrix}$$

$P^{(1)}$称为马尔可夫链的一阶转移概率矩阵。它的元素为非负的，且行元素之和等于1。

如果过程的状态不是三种，而是有 m 种，即 $E_1,E_2,\cdots E_m$，那么由状态 E_i 经过一步转移到状态 E_j 的一阶转移概率矩阵为：

$$P^{(1)}=\begin{pmatrix}p_{11}^{(1)} & p_{12}^{(1)} & \cdots & p_{1m}^{(1)}\\ p_{21}^{(1)} & p_{22}^{(1)} & \cdots & p_{2m}^{(1)}\\ \vdots & \vdots & \vdots & \vdots\\ p_{m1}^{(1)} & p_{m2}^{(1)} & \cdots & p_{mm}^{(1)}\end{pmatrix}$$

由条件分布律的性质，知转移概率有如下性质：

$$0\leqslant p_{ij}^{(1)}\leqslant 1$$

$$\sum_{j=1}^{m}p_{ij}^{(1)}=1 \qquad (i=1,2,\cdots,m)$$

二、高阶转移概率

如果马尔可夫链有 m 种状态 $E_1,E_2,\cdots,E_m$，从状态 E_i 出发经两步转移到状态 E_j 的概率（不管第一步是什么状态）称为二阶转移概率，记为 $p_{ij}^{(2)}$，用二阶转移概率排成的矩阵为

$$P^{(2)}=[p_{ij}^{(2)}]_{m\times m}=\begin{pmatrix}p_{11}^{(2)} & p_{12}^{(2)} & \cdots & p_{1m}^{(2)}\\ p_{21}^{(2)} & p_{22}^{(2)} & \cdots & p_{2m}^{(2)}\\ \vdots & \vdots & \vdots & \vdots\\ p_{m1}^{(2)} & p_{m2}^{(2)} & \cdots & p_{mm}^{(2)}\end{pmatrix}$$

这个矩阵称为二阶转移概率矩阵，其中元素 $p_{ij}^{(2)}$ 可以由实际资料统计出来，即

$$p_{ij}^{(2)}=\frac{E_i\text{ 后的第二步是 }E_j\text{ 的次数}}{E_i\text{ 出现的次数}}$$

更一般地，由状态 E_i 经 k 步转移到状态 E_j 的概率 $p_{ij}^{(k)}$ 称为 k 阶转移概率，其转移概率矩阵

$$P^{(k)}=[p_{ij}^{(k)}]_{m\times m}=\begin{pmatrix}p_{11}^{(k)} & p_{12}^{(k)} & \cdots & p_{1m}^{(k)}\\ p_{21}^{(k)} & p_{22}^{(k)} & \cdots & p_{2m}^{(k)}\\ \vdots & \vdots & \vdots & \vdots\\ p_{m1}^{(k)} & p_{m2}^{(k)} & \cdots & p_{mm}^{(k)}\end{pmatrix}$$

称为 k 阶转移概率矩阵，其中

$$p_{ij}^{(k)}=\frac{E_i\text{ 后的第 }k\text{ 步是 }E_j\text{ 的次数}}{E_i\text{ 出现的次数}}$$

且有性质 $0\leqslant p_{ij}^{(k)}\leqslant 1$；$\sum_{j=1}^{m}p_{ij}^{(k)}=1$。

对于高阶转移概率的计算，事实上可以利用一阶转移概率根据马尔可夫链的无后效性而得到，例如，对于二阶转移概率

$$\begin{aligned}p_{ij}^{(2)}&=p\{x_2=j|x_0=i\}\\&=p\{x_1=1,x_2=j|x_0=i\}+p\{x_1=2,x_2=j|x_0=i\}+\cdots\end{aligned}$$

$$+p\{x_1=m,x_2=j|x_0=i\}$$

$$=\frac{p\{x_0=i,x_1=1,x_2=j\}}{p\{x_0=i\}}+\frac{p\{x_0=i,x_1=2,x_2=j\}}{p\{x_0=i\}}+\cdots$$

$$+\frac{p\{x_0=i,x_1=m,x_2=j\}}{p\{x_0=i\}}$$

$$=\frac{p\{x_0=i,x_1=1\}}{p\{x_0=i\}}p\{x_2=j|x_0=i,x_1=1\}$$

$$+\frac{p\{x_0=i,x_1=2\}}{p\{x_0=i\}}p\{x_2=j|x_0=i,x_1=2\}+\cdots$$

$$+\frac{p\{x_0=i,x_1=m\}}{p\{x_0=i\}}p\{x_2=j|x_0=i,x_1=m\}$$

$$=p_{i1}^{(1)}p_{1j}^{(1)}+p_{i2}^{(1)}p_{2j}^{(1)}+\cdots+p_{im}^{(1)}p_{mj}^{(1)}=\sum_{k=1}^{m}p_{ik}^{(1)}p_{kj}^{(1)} \tag{9-6}$$

因而有

$$P^{(2)}=\begin{bmatrix} p_{11}^{(2)} & p_{12}^{(2)} & \cdots & p_{1m}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} & \cdots & p_{2m}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(2)} & p_{m2}^{(2)} & \cdots & p_{mm}^{(2)} \end{bmatrix}$$

$$=\begin{bmatrix} \sum_{k=1}^{m}p_{1k}^{(1)}p_{k1}^{(1)} & \sum_{k=1}^{m}p_{1k}^{(1)}p_{k2}^{(1)} & \cdots & \sum_{k=1}^{m}p_{1k}^{(1)}p_{km}^{(1)} \\ \sum_{k=1}^{m}p_{2k}^{(1)}p_{k1}^{(1)} & \sum_{k=1}^{m}p_{2k}^{(1)}p_{k2}^{(1)} & \cdots & \sum_{k=1}^{m}p_{2k}^{(1)}p_{km}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{k=1}^{m}p_{mk}^{(1)}p_{k1}^{(1)} & \sum_{k=1}^{m}p_{mk}^{(1)}p_{k2}^{(1)} & \cdots & \sum_{k=1}^{m}p_{mk}^{(1)}p_{km}^{(1)} \end{bmatrix}$$

$$=\begin{bmatrix} p_{11}^{(1)} & p_{12}^{(1)} & \cdots & p_{1m}^{(1)} \\ p_{21}^{(1)} & p_{22}^{(1)} & \cdots & p_{2m}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(1)} & p_{m2}^{(1)} & \cdots & p_{mm}^{(1)} \end{bmatrix}\begin{bmatrix} p_{11}^{(1)} & p_{12}^{(1)} & \cdots & p_{1m}^{(1)} \\ p_{21}^{(1)} & p_{22}^{(1)} & \cdots & p_{2m}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(1)} & p_{m2}^{(1)} & \cdots & p_{mm}^{(1)} \end{bmatrix}$$

$$=P^{(1)}\cdot P^{(1)}=(P^{(1)})^2$$

从而，对于高阶转移概率矩阵有

$$P^{(k)}=(P^{(1)})^k$$

更一般的情况是，对于任何 r，可以导出

$$p_{ij}^{(k)}=\sum_{l=1}^{m}p_{il}^{(r)}p_{lj}^{(k-r)} \tag{9-7}$$

也就是说，从状态 E_i 出发经过 k 步到达状态 E_j 这一过程，可以看作它是先经过 $r(0<r<k)$ 步转移到某一状态 $E_l(l=1,2,\cdots,m)$，再由 E_l 经过 $(k-r)$ 步转移到达状态 E_j。

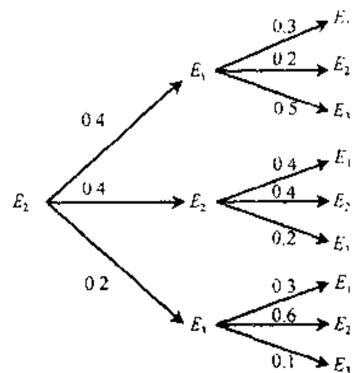


图 9-1　三种状态二阶转移概率示意图

例如，对于一个包含砂岩 (E_1)、粉砂岩 (E_2) 和页岩 E_3 的剖面，由图 9-1 可见，从 E_2 出发经过两步转移到 E_3 有三条不同的途径，根据式(9-7)，则有

$$p_{23}^{(2)} = \sum_{j=1}^{3} p_{2j}^{(1)} p_{j3}^{(1)} = p_{21}^{(1)} p_{13}^{(1)} + p_{22}^{(1)} p_{23}^{(1)} + p_{23}^{(1)} p_{33}^{(1)}$$
$$= 0.4 \times 0.5 + 0.4 \times 0.2 + 0.2 \times 0.1 = 0.30$$

它是从 E_2 出发由三条不同途径经过两步转移到 E_3 的概率之和。同样可以计算：

$$p_{22}^{(2)} = \sum_{j=1}^{3} p_{2j}^{(1)} p_{j2}^{(1)} = p_{21}^{(1)} p_{12}^{(1)} + p_{22}^{(1)} p_{22}^{(1)} + p_{23}^{(1)} p_{32}^{(1)}$$
$$= 0.4 \times 0.2 + 0.4 \times 0.4 + 0.2 \times 0.6 = 0.36$$

$$p_{21}^{(2)} = \sum_{j=1}^{3} p_{2j}^{(1)} p_{j1}^{(1)} = p_{21}^{(1)} p_{11}^{(1)} + p_{22}^{(1)} p_{21}^{(1)} + p_{23}^{(1)} p_{31}^{(1)}$$
$$= 0.4 \times 0.3 + 0.4 \times 0.4 + 0.2 \times 0.3 = 0.34。$$

§3 遍历定理与极限分布

马尔可夫链遍历性的直观意义是：不论从哪个初始状态 E_i 出发，当转移步数 k 充分大后，它到达状态 E_j 的概率是一个不随时间变化的常数 p_j。也就是说，无论初始状态如何，经过若干步转移以后，系统将处于平衡状态，因而当 k 充分大时，可用 p_j 作为 $p_{ij}^{(k)}$ 的近似值。这样，便可以解决当 k 很大时高阶转移概率的计算问题。p_j 称为马尔可夫链的极限概率，而遍历性的中心问题是要确定在什么样的条件下，转移概率的极限才是存在的；极限概率是否构成一个概率分布；以及如何计算极限概率 p_j。

遍历性定理是指对于有限状态的马尔可夫链，若存在一个正整数 s，使得 $p_{ij}^{(s)}>0$ 对任何 i，$j=1,2,\cdots,m$ 成立，那么极限

$$\lim_{n\to\infty} p_{ij}^{(n)} = p_j \tag{9-8}$$

存在，并且与 i 无关；而式(9-8)中的 $\{p_1,p_2,\cdots,p_m\}$ 是方程组

$$p_j = \sum_{i=1}^{m} p_i p_{ij}^{(1)} \quad (j = 1,2,\cdots,m)$$

在满足条件 $p_j>0$，$\sum_{i=1}^{m} p_j=1$ 时的唯一解。

例如，有一马尔可夫链，其转移状态有两种：E_1，E_2。经计算得出它的一阶转移概率矩阵为

$$P^{(1)} = \begin{bmatrix} 0.79 & 0.21 \\ 0.59 & 0.41 \end{bmatrix}$$

当 $s=1$ 时，对一切 i,j，$p_{ij}^{(1)}>0$ 满足遍历性定理，故有 $\lim_{n\to\infty} p_{ij}^{(n)}=p_j>0$。而 p_j 可由方程组

$$\begin{cases} p_j = \sum_{i}^{m} p_i p_{ij}^{(1)} & (j = 1,2,\cdots,m) \\ \sum_{j=1}^{m} p_j = 1 & (p_j > 0) \end{cases}$$

求出。对于本例为

$$\begin{cases} p_1 = 0.79p_1 + 0.59p_2 \\ p_2 = 0.21p_1 + 0.41p_2 \\ p_1 + p_2 = 1 & (p_1, p_2 > 0) \end{cases}$$

最后得到 $p_2=0.26$，$p_1=0.74$。所以，其极限概率矩阵为

$$\tilde{p}=\begin{bmatrix}0.74 & 0.26\\0.74 & 0.26\end{bmatrix}$$

如果从公式 $P^{(k)}=(P^{(1)})^k$ 出发,计算其高阶转移概率有

$$P^{(2)}=P^{(1)}\cdot P^{(1)}=\begin{bmatrix}0.75 & 0.25\\0.71 & 0.29\end{bmatrix}$$

$$P^{(3)}=P^{(2)}\cdot P^{(1)}=\begin{bmatrix}0.74 & 0.26\\0.74 & 0.26\end{bmatrix}$$

$$P^{(4)}=P^{(3)}\cdot P^{(1)}=\begin{bmatrix}0.74 & 0.26\\0.74 & 0.26\end{bmatrix}$$

从各阶转移概率可以看出,其前三阶有所不同,随着阶数增加,3、4 阶转移概率矩阵相等,等于极限概率矩阵,而且矩阵中每一列内各元素均相等,即经过若干步转移后,终止状态 E_j 的概率是一个常数 p_j。这就是状态 E_j 的极限概率。

§4 马尔可夫概型检验

对任何一个离散的时间序列或空间序列都可以构造出一个转移概率矩阵。但是,它是否具有马尔可夫概型的性质呢?这就要对其进行独立性检验。通常是用 χ^2 检验。

皮尔逊已经证明统计量

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{m}\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2\Big/\frac{n_{i.}n_{.j}}{n} \tag{9-9}$$

在独立假设下,当 n 很大时服从自由度 $(m-1)^2$ 的 χ^2 分布。其中 m 为过程状态的种数;n_{ij}为转移频数(由状态 E_i 经过一步转移到 E_j 上的次数)。

$$n_{i.}=\sum_{j=1}^{m}n_{ij};\quad n_{.j}=\sum_{i=1}^{m}n_{ij}$$

$$n=\sum_{i=1}^{m}n_{i.}=\sum_{j=1}^{m}n_{.j}$$

例如,设有一马尔可夫链,其转移状态有两种:E_1、E_2。经统计得转移频数,见表 9-1。

假设 H_0:x_t 与 x_{t1}是独立的。为检验这一假设,计算统计量

表 9-1 状态转移频数

x_{t1}	x_t		
	n_{ij}		$n_{i\cdot}$
	E_1	E_2	
E_1	7	3	10
E_2	3	2	5
$n_{\cdot j}$	10	5	=15

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\left(n_{ij}-\frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2\Big/\frac{n_{i\cdot}n_{\cdot j}}{n}$$

$$=\frac{\left(n_{11}-\frac{n_{1\cdot}n_{\cdot 1}}{n}\right)^2}{\frac{n_{1\cdot}n_{\cdot 1}}{n}}+\frac{\left(n_{12}-\frac{n_{1\cdot}n_{\cdot 2}}{n}\right)^2}{\frac{n_{1\cdot}n_{\cdot 2}}{n}}+\frac{\left(n_{21}-\frac{n_{2\cdot}n_{\cdot 1}}{n}\right)^2}{\frac{n_{2\cdot}n_{\cdot 1}}{n}}+\frac{\left(n_{22}-\frac{n_{2\cdot}n_{\cdot 2}}{n}\right)^2}{\frac{n_{2\cdot}n_{\cdot 2}}{n}}$$

$$=\frac{\left(7-\frac{10\times 10}{15}\right)^2}{\frac{10\times 10}{15}}+\frac{\left(3-\frac{10\times 5}{15}\right)^2}{\frac{10\times 5}{15}}+\frac{\left(3-\frac{5\times 10}{15}\right)^2}{\frac{5\times 10}{15}}+\frac{\left(2-\frac{5\times 5}{15}\right)^2}{\frac{5\times 5}{15}}$$

由于统计量 χ^2 只是在 n 很大时服从自由度为 $(m-1)$ 的 χ^2 分布，而上例中 $m=2$，即自由度为 1，必须修正 $n_{i\cdot}n_{\cdot j}/n$ 的值，耶斯定为 0.5，即

$$\chi^2=\frac{\left(7-\frac{20}{3}-0.5\right)^2}{\frac{20}{3}}+\frac{\left(3-\frac{10}{3}-0.5\right)^2}{\frac{10}{3}}+\frac{\left(3-\frac{10}{3}-0.5\right)^2}{\frac{10}{3}}+\frac{\left(2-\frac{5}{3}-0.5\right)^2}{\frac{5}{3}}$$

$$=0.438$$

而 $(m-1)^2=(2-1)^2=1$，自由度为 1 时，$\chi^2_{0.05}=3.841$

因 $\chi^2=0.438<3.841$，故接受 H_0 假设，认为 t_1 时刻过程处于什么状态与 t 时刻过程所处状态无关，故该过程并非马尔可夫链。

§5 应用算例

【例 1】 马尔可夫链一阶转移概率矩阵研究沉积回旋的简例(据成都地质学院)：剖面采自某地区钻孔资料，岩性为紫红色粉砂质泥岩夹长石砂岩透镜体，地层层数共 90 层($n=90$)；状态分为五种($m=5$)：E_1(砂砾岩)、E_2(粉砂岩)、E_3(泥、页、菱铁泥岩)、E_4(根土、砂、粉砂页岩)、E_5(煤、炭质页岩)。

1. 转移频数矩阵

$$Q=[n_{ij}]=\begin{array}{c} \\ E_1\\E_2\\E_3\\E_4\\E_5\\n_{\cdot j}\end{array}\begin{array}{c}\begin{array}{ccccc}E_1&E_2&E_3&E_4&E_5\end{array}\\ \begin{bmatrix}0&7&2&6&2\\7&0&2&12&1\\2&1&0&1&3\\1&6&1&0&15\\6&8&2&4&0\end{bmatrix}\\ \begin{array}{ccccc}16&22&7&23&21\end{array}\end{array}\begin{array}{c}n_{i\cdot}\\17\\22\\7\\23\\20\\89\end{array}$$

2. 由 n_{ij} 可以求得频率转移概率 $\hat{P}_{ij}=\frac{n_{ij}}{n_{i\cdot}}$，用它作为理论概率估计值而得一阶转移概率矩阵为

$$P^{(1)} = [\hat{p}_{ij}] = \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array} \begin{array}{c} \begin{array}{ccccc} E_1 & E_2 & E_3 & E_4 & E_5 \end{array} \\ \begin{bmatrix} 0 & 0.41 & 0.12 & 0.35 & 0.12 \\ 0.32 & 0 & 0.09 & 0.54 & 0.04 \\ 0.28 & 0.14 & 0 & 0.14 & 0.43 \\ 0.04 & 0.26 & 0.04 & 0 & 0.65 \\ 0.30 & 0.40 & 0.10 & 0.20 & 0 \end{bmatrix} \end{array}$$

3. 由 $P^{(1)}$ 求其极限概率而得

$$\widetilde{P} = \begin{array}{c} \\ E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \end{array} \begin{array}{c} \begin{array}{ccccc} E_1 & E_2 & E_3 & E_4 & E_5 \end{array} \\ \begin{bmatrix} 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \end{bmatrix} \end{array}$$

4. 简化旋回模式

先取出固定向量为[0.1826、0.2471、0.0786、0.2576、0.2340]。由其中具有最大概率的状态作为旋回的开始,故我们可取 E_4 作为开始,直接利用 $P^{(1)}$ 来画出旋回模式,如图 9-2。

如果加以简化,可以得到主要旋回模式为

$$E_4 \to E_5 \to E_2 \to E_4$$

或

$$E_4 \to E_5 \to E_1 \to E_2 \to E_4$$

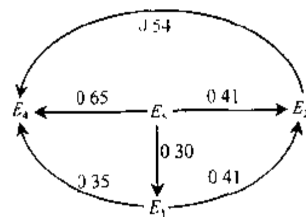


图 9-2 旋回模式

【例 2】 马尔可夫链在航空照片水系研究中的应用(据长春地质学院):根据水系分布特点结合地质条件,对某水系集结区按水系模型中水道的方向不同划分为六种状态,并编号为:SN(1)、NNE(2)、NE(3)、EW(4)、NW(5)、NNW(6)。水系模型可分为一级、二级、三级和四级水道系统。把水道的主流(一级)和支流(二、三、四……级)的方向性变化关系看成一种状态向另一种状态的转移。可以认为这种状态的转移是随机的,具有马尔可夫链的性质。

1. 某水系分布范围水道方向转移频数表(表 9-2)

2. 由 $Q=[n_{ij}]$ 计算一阶转移概率矩阵

$$P^{(1)} = [\hat{p}_{ij}] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 \\ 0 & 0.056 & 0.141 & 0.560 & 0.028 & 0.211 \\ 0.083 & 0.417 & 0.166 & 0.166 & 0.041 & 0.125 \\ 0.137 & 0.096 & 0.370 & 0.096 & 0.219 & 0.082 \\ 0.064 & 0.234 & 0.300 & 0.277 & 0.021 & 0.106 \\ 0 & 0.071 & 0.572 & 0.143 & 0 & 0.214 \end{bmatrix} \end{array}$$

表 9-2　水道方向转移频数

状态		SN	NNE	NE	EW	NW	NNW	$n_{i\cdot}$
		1	2	3	4	5	6	
		n_{ij}						
SN	1	1	1	1	1	1	1	6
NNE	2	0	4	10	40	2	15	71
NE	3	2	10	4	4	1	3	24
EW	4	10	7	27	7	16	6	73
NW	5	3	11	14	13	1	5	47
NNW	6	0	1	8	2	0	3	14
$n_{\cdot j}$		16	34	64	67	21	33	235

3. 由 $P^{(1)}$ 求其极限概率矩阵

$$\tilde{P}=\begin{array}{c} \\ 1\\2\\3\\4\\5\\6\end{array}\begin{array}{c}\begin{array}{cccccc}1 & 2 & 3 & 4 & 5 & 6\end{array}\\ \begin{bmatrix}0.072 & 0.190 & 0.279 & 0.231 & 0.081 & 0.146\\ 0.072 & 0.190 & 0.279 & 0.231 & 0.081 & 0.146\\ 0.072 & 0.190 & 0.279 & 0.231 & 0.081 & 0.146\\ 0.072 & 0.190 & 0.279 & 0.231 & 0.081 & 0.146\\ 0.072 & 0.190 & 0.279 & 0.231 & 0.081 & 0.146\\ 0.072 & 0.190 & 0.279 & 0.231 & 0.081 & 0.146\end{bmatrix}\end{array}$$

由水道方向转移频数表可见，水道转移频数较大组为EW(4)、NE(3)、NNE(2)三组。这和极限概率较大组基本一致，即水道较多的方向是EW(4)、NE(3)、NNE(2)三组。这表明了极限概率基本上反映了水系形成的特点。

【例 3】 马尔可夫链在地层柱状剖面火山喷发旋回中的应用(据长春地质学院)。

1. 问题

石炭系下统鹿圈屯组为一富钠质的细碧角斑岩建造，由下而上划分为两个旋回。第一旋回由石英角斑岩类、凝灰质砂岩类、细碧角斑岩类、泥沙质岩或碳酸盐类岩石组成；第二旋回以石英斑岩类(或英安岩类)、泥砂质岩和碳酸盐岩类为组合特征。第一旋回和第二旋回之间是以碳酸盐类岩石为标志层分开。第一旋回顶为明显标志层(大理岩)，形成一个有顶无底的地层，再加上半覆盖区露头稀少的客观条件，限制了给第一旋回地层的进一步划分，给对比工作带来一定的困难。

采用马尔可夫链方法对石炭系下统鹿圈屯组第一旋回变质火山岩系进行地层模拟工作，以对实测地层柱状剖面的对比和地层旋回的正确划分。

2. 系统状态

根据地层实测剖面资料将那些外表特征略有差异而本质岩性相近的岩层进行岩层类型的划分和统一编号，把地层剖面的不同岩性的岩层设想为一个系统中的若干状态，共划分为九种状态即：

E_1(细碧玢岩含角砾的凝灰熔岩)、E_2(石英角斑岩)、E_3(细碧玢岩凝灰岩)、E_4(角斑质凝灰岩)、E_5(细碧玢岩)、E_6(细碧玢岩凝灰熔岩)、E_7(凝灰质砂岩)、E_8(千枚状板岩)、E_9(大理岩)。

3. 转移频数矩阵表

从实测地层柱状图上由底部向上部按设计的岩层状态编号编制转移频数表，地层柱状图的不同岩层的变化可以看成马尔可夫链状态的转移，由于不同实测剖面所出现的岩层状态是不完全相同的，故它们所形成的岩石状态可以大同小异。如第Ⅰ剖面的岩石状态为 E_1、E_2、E_3、E_4、E_6、E_7 六种状态；第Ⅱ剖面的岩层状态为 E_1、E_2、E_4、E_5、E_7、E_9 六种状态。由两实测剖面可分别作出它们的转移频数矩阵表表 9-3 和表 9-4

表 9-3 剖面Ⅰ岩层转移频数矩阵表

状态	状态						
	E_1	E_2	E_3	E_4	E_6	E_6	行和($n_{i\cdot}$)
	n_{ij}						
E_1	0	0	0	1	0	0	1
E_2	0	0	1	3	0	1	5
E_3	1	1	0	2	0	0	4
E_4	0	3	3	0	3	0	9
E_6	0	0	0	3	0	0	3
E_7	0	0	0	1	0	0	1
列和($n_{\cdot j}$)	1	4	4	10	3	1	23

表 9-4 剖面Ⅱ岩层转移频数矩阵表

状态	状态						
	E_1	E_2	E_3	E_4	E_6	E_6	行和($n_{i\cdot}$)
	n_{ij}						
E_1	0	0	0	1	0	0	1
E_2	0	0	1	0	1	0	2
E_4	1	0	0	3	0	1	5
E_5	0	1	4	0	0	0	5
E_7	0	0	0	1	0	0	1
E_9	0	0	1	0	0	0	1
列和($n_{\cdot j}$)	1	1	6	5	1	1	15

4. 转移频数矩阵和极限概率

由表 9-3，表 9-4 分别计算其一阶转移概率矩阵 $P_{\mathrm{I}}^{(1)}$、$P_{\mathrm{II}}^{(1)}$ 和它们的极限概率 $\tilde{P}_{\mathrm{I}}$、$\tilde{P}_{\mathrm{II}}$：

$$
\begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 6 & 7 \end{array} \\
P_{\mathrm{I}}^{(1)} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 6 \\ 7 \end{array} & \left[\begin{array}{cccccc} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.20 & 0.60 & 0 & 0.20 \\ 0.25 & 0.25 & 0 & 0.50 & 0 & 0 \\ 0 & 0.33 & 0.33 & 0 & 0.33 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{array}\right] \end{array}
$$

$$\widetilde{P}_{t}=\begin{bmatrix}1 & 2 & 3 & 4 & 5 & 6\\ 0.04 & 0.18 & 0.18 & 0.42 & 0.14 & 0.04\end{bmatrix}$$

$$P_{1}^{(1)}=\begin{array}{c} \\ 1\\2\\4\\5\\7\\9\end{array}\begin{array}{c}\begin{array}{cccccc}1 & 2 & 4 & 5 & 7 & 9\end{array}\\ \begin{bmatrix}0 & 0 & 0 & 1.0 & 0 & 0\\ 0 & 0 & 0.50 & 0 & 0.50 & 0\\ 0.20 & 0 & 0 & 0.60 & 0 & 0.20\\ 0 & 0.20 & 0.80 & 0 & 0 & 0\\ 0 & 0 & 0 & 1.0 & 0 & 0\\ 0 & 0 & 1.0 & 0 & 0 & 0\end{bmatrix}\end{array}$$

$$\widetilde{P}_{t}=\begin{bmatrix}1 & 2 & 4 & 5 & 7 & 9\\ 0.08 & 0.07 & 0.39 & 0.35 & 0.03 & 0.08\end{bmatrix}$$

5. 成果解释

从Ⅰ、Ⅱ两地层柱状剖面的转移频数，转移概率和极限概率可以看出，第四种状态(角斑质凝灰岩)的转移频数，转移概率和极限概率均为最大，这一事实表明，Ⅰ、Ⅱ两剖面主要岩石类型在数量方面存在着明显的相似性。

在极限概率的基础上可作出两剖面岩层的模拟柱状图，模拟剖面Ⅰ明显地反映出火山喷发旋回的特点，每个旋回以角斑凝灰岩(E_4)开始，随后是细碧玢岩凝灰岩(E_3)或细碧玢岩凝灰熔岩(E_6)的基性岩喷发；模拟剖面Ⅱ明显的火山喷发旋回特点是每个喷发旋回以角斑质凝灰岩、角斑质凝灰熔岩(E_4)开始，相继是深灰色细碧玢岩(E_5)，随之大理岩(E_9)的正常沉积。

通过模拟剖面旋回规律展示，再同实测地层柱状剖面对比，它有助于对实测地层柱状剖面的地层旋回正确划分。Ⅰ剖面和模拟地层柱状剖面对比可以划分为10个旋回；Ⅱ剖面和模拟地层柱状剖面对比可以划分7个旋回，通过两个实测地层柱状剖面的旋回进一步划分，为两个实测地层柱状图的旋回对比提供了地质依据。它们的喷发旋回次数和各自旋回岩相的组合特点是有一定差异的。但它们有的共同特点是由角斑质凝灰岩、凝灰熔岩(E_4)开始，相继喷发的是细碧玢岩凝灰岩(E_3)或者是细碧玢岩凝灰熔岩(E_6)、深灰色细碧玢岩(E_5)，即喷发旋回表现为由酸性到基性的特点。因此，这两个地层柱状剖面的火山喷发旋回是可以对比的。

【例4】 马尔可夫链分析在新疆拜城亚格列木组下段沉积相研究中的应用(摘自成都地质学院学报第15卷，何开华)。

亚格列木组(k_{1y})属下白垩统，其剖面位于新疆拜城县北约50km的卡普沙河谷东岸，露头良好，k_{1y}下段厚约155m，为一套山麓相的棕红色细—中砾岩和砂岩，沉积环境主要为冲积扇和辫状河。

不考虑厚度，只以各岩性或岩相的向上转移作为观察计数的标准。这样计数的地质意义是：同一种岩类，即使本身的厚度再大，都是沉积补偿作用下形成同一环境的产物。只有不同岩性或岩相类型的转移，才反映存在的地质背景的差异。因此该取样法优点就是确切地表达了沉积作用演化的信息，突出了沉积作用的相序列，有利于环境的划分和分析。

根据岩性、粒度和沉积构造因素，将k_{1y}下段分为七个微岩相：(1)冲刷相、(2)滞留砾岩相、(3)含砾砂岩相、(4)板状交错层理砂岩相、(5)槽状交错层理砂岩相、(6)平行层理砂岩相、(7)水平层理砂岩相。由微岩相演化序列的柱状部面图统计出自下而上的转移频数矩阵如

下：

$$
Q = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array}
\begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{bmatrix}
0 & 14 & 2 & 0 & 0 & 0 & 0 \\
0 & 0 & 12 & 0 & 0 & 0 & 2 \\
2 & 0 & 0 & 10 & 2 & 1 & 10 \\
5 & 0 & 1 & 0 & 9 & 2 & 4 \\
2 & 0 & 2 & 1 & 0 & 1 & 5 \\
2 & 0 & 1 & 4 & 0 & 0 & 0 \\
5 & 0 & 6 & 6 & 1 & 3 & 0
\end{bmatrix} \end{array}
$$

由转移频数矩阵，根据式(9-9)算出统计量 $\chi^2=180.17$，此时的自由度为$(7-1)^2=36$，查χ^2分布表，当取 $a=0.05$ 时的临界值小于 60，故 $\chi^2>\chi^2_{0.05}$，因而该序列为一马尔可夫链。

从转移频数矩阵 Q 可求得转移概率矩阵为

$$
P = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array}
\begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\
\begin{bmatrix}
0 & 0.875 & 0.125 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.857 & 0 & 0 & 0 & 0.143 \\
0.08 & 0 & 0 & 0.40 & 0.08 & 0.04 & 0.4 \\
0.238 & 0 & 0.048 & 0 & 0.429 & 0.095 & 0.19 \\
0.182 & 0 & 0.182 & 0.091 & 0 & 0.091 & 0.455 \\
0.286 & 0 & 0.143 & 0.571 & 0 & 0 & 0 \\
0.238 & 0 & 0.286 & 0.286 & 0.048 & 0.143 & 0
\end{bmatrix} \end{array}
$$

在岩性的转移是“随机的”这一假设下，可以求出随机序列转移概率频数 $R=[r_{ij}]$，其中 $r_{ij}=n_{\cdot j}/(n-n_{i\cdot})$。而 P—R 称做差值矩阵 $D=[d_{ij}]$，它的意义是：如果 $d_{ij}>0$ 意味着观察到的实际岩性转移比随机岩性转移常见，$d_{ij}<0$ 意味着观察的岩性转移比随机岩性转移少见。于是，就可依据差值矩阵 D 中的正值元素，作出岩相旋回进程线路图，并最后作出沉积序列模式。今算得：

$$
R = \begin{bmatrix}
0 & 0.141 & 0.242 & 0.212 & 0.121 & 0.071 & 0.212 \\
0.158 & 0 & 0.238 & 0.108 & 0.119 & 0.069 & 0.208 \\
0.178 & 0.156 & 0 & 0.233 & 0.133 & 0.078 & 0.233 \\
0.170 & 0.149 & 0.255 & 0 & 0.128 & 0.074 & 0.223 \\
0.154 & 0.135 & 0.231 & 0.202 & 0 & 0.067 & 0.202 \\
0.148 & 0.130 & 0.222 & 0.194 & 0.111 & 0 & 0.194 \\
0.170 & 0.149 & 0.255 & 0.223 & 0.128 & 0.074 & 0
\end{bmatrix}
$$

$$D=\begin{pmatrix} 0 & 0.734 & -0.117 & -0.212 & -0.121 & -0.071 & -0.212 \\ -0.158 & 0 & 0.620 & -0.208 & -0.119 & -0.069 & -0.065 \\ -0.098 & -0.156 & 0 & 0.167 & -0.053 & -0.038 & 0.167 \\ 0.068 & -0.149 & -0.208 & 0 & 0.301 & 0.021 & -0.033 \\ 0.028 & -0.135 & -0.049 & -0.111 & 0 & 0.024 & 0.253 \\ 0.138 & -0.130 & -0.079 & 0.377 & -0.111 & 0 & -0.194 \\ 0.068 & -0.149 & 0.030 & 0.062 & -0.08 & 0.068 & 0 \end{pmatrix}$$

从 D 中挑出其正值元素，并按每行正值的大小顺序、排列成下表：

(1,2)=0.734			
(2,3)=0.620			
(3,4)=0.167	(3,7)=0.167		
(4,5)=0.301	(4,1)=0.068	(4,6)=0.021	
(5,7)=0.253	(5,1)=0.028	(5,6)=0.024	
(6,4)=0.377	(6,1)=0.138		
(7,6)=0.068	(7,1)=0.068	(7,4)=0.062	(7,3)=0.03

由于观察时的某些不利因素（如覆盖、断裂等等），往往对转移不能做到全面的统计，故并非所有正值元素都反映了客观实际，选用那一些，经验的作法是：(1) 确定一个下限标准，低于这个标准的不选用；(2) 个别大的正值元素，在作岩相的旋回进程线路图中，在线路联接上反映出有地质解释上难于接受的干扰时，也要加以舍弃。

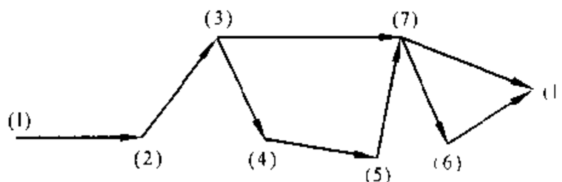


图 9-3 k_{1y}下段各微岩相的旋回进程线路图

现下限标准定为 0.068，正值元素表中的 16 个正值元素，低于标准的有 5 个，其余 11 个因线路联结不合理放弃的有 2 个，它们是(6,4)和(4,1)，共选用了 9 个正值元素，占全部选用元素的 9/11＝82%，选用率是较高的。

以选取的 9 个正值元素为依据，绘出旋回进程线路图，如图 9-3 所示。在此基础上绘制出 k_{1y}下段的沉积序列模式图，如图 9-4 所示。该模式图较清楚地反映了该地区早白垩世初期冲积扇的辫状河微岩相演化发展的特征。

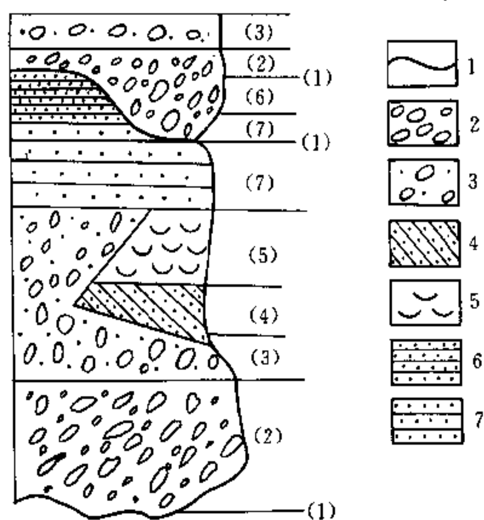


图 9-4 沉积序列模式图

模式剖面的下部，表明辫状河初始期具有槽洪（stream flood）的搬运和沉积作用的性质，碎屑流和牵引流混合存在，流势猛烈，泥砂夹带砾石，顺坡下泻，既使早先的坡面经受冲刷，又在适当处所沉积了基

质含量颇高的滞流砾石和河道砾石。之后，流速虽然未减，但在流量已逐渐减小的情况下，沉积了含砾砂岩。由于大量粗碎屑快速地分离出来，使得流体密度大为降低，于是形成了较为正常的牵引流。这时含砂砾岩便可过渡为发育不好的边滩沉积；也可过渡为分流的河间砂坝。由于辫状河流经的山麓地带，地形坡度减小不多，所以辫状河的弯曲度都不大(求得的弯曲度指数为 1.4—1.7)，这使得辫状河仍保持较大的流速，出现了上部水流动态，形成了休止角小的交错层理和平行层理。这以后又开始新的暴期的到来，转化为另一个相类似的沉积旋回序列。

习　题

1. 什么是马尔可夫链？什么是马尔可夫链一阶转移概率？什么是马尔可夫链的 l 阶转移概率？什么情况下称马尔可夫链是平稳的？
2. 如何统计一个地层剖面的岩性转移频数，得出转移频数矩阵？如何求它的转移频率(概率)矩阵？
3. 什么样的马尔可夫链具有极限概率？
4. 某地层剖面由(1)亮晶异常化学岩、(2)微晶异常化学岩、(3)微晶石灰岩、(4)白云岩等四种岩性所组成，今统计得出其转移频数矩阵为：

$$\begin{array}{c} \\ (1) \\ (2) \\ (3) \\ (4) \end{array} \begin{array}{c} \begin{array}{cccc} (1) & (2) & (3) & (4) \end{array} \\ \begin{bmatrix} 0 & 7 & 7 & 3 \\ 6 & 0 & 5 & 1 \\ 8 & 4 & 0 & 4 \\ 2 & 1 & 5 & 0 \end{bmatrix} \end{array}$$

① 试检验该岩性转移序列是否具有马尔可夫链概型性质，(取 $a=0.05$)；
② 求岩性转移概率矩阵 P；
③ 求随机转移概率矩阵 R；
④ 求差值矩阵 D，并挑出正值元素，按每行正值的大小列表；
⑤ 试根据正值元素作出岩相的旋回进程线路图。

第十章　蒙特卡罗法

§1　蒙特卡罗法概述

统计试验法又名蒙特卡罗(Monte Carlo)法。它的基本思想可以概括为:欲求给定问题的数值解,则先构造一个表征给定问题的概率模型

$$Y = \mu(X_1, X_2, \cdots, X_n) \tag{10-1}$$

使得要求的数值解恰好是概率模型式(10-1)的某个数字特征(如数学期望等),而这个数字特征又可用统计的方法求得其估计值,我们就把这个估计值作为给定问题的近似值。

由蒙特卡罗法的基本思想可以得出:

① 蒙特卡罗法求给定问题数值解的过程大致可归结为四步,即:构造给定问题的概率模型;对概率模型中的随机变量 $X_1, X_2, \cdots, X_n$ 进行观测(抽样),获得随机变量的若干组观测值 $x_{1i}, x_{2i}, \cdots, x_{ni}(i=1,2,\cdots)$;根据随机变量的观测值由概率模型式(10-1)求得随机变量 Y 的一系列估计算值 $y_1, y_2, \cdots$;最后用统计的方法由 y_1、y_2,…求出给定问题的近似解。

在上述求解的过程中,关键的一步是对随机变量的随机抽样,这个问题的解决有赖于下一节介绍的随机数。

② 只要构造出了给定问题的概率模型,并解决了对模型中随机变量的抽样问题,就可用统计的方法将概率模型的数字特征估计出来,因此,蒙特卡罗法并非是对某特定问题求解的特有方法,它是以概率论与数理统计理论为指导的有着广泛应用领域的通用性统计学方法。

这种统计试验方法需要对模型中不同分布的随机变量进行大量的抽样计算,工作量太大,很多问题的求解靠人工几乎不能实现。本世纪 40 年代电子计算机的问世,解决了计算工具问题,才有可能实现大量的随机抽样试验,并用蒙特卡罗法来研究和解决实际问题。当时最具代表性的实际应用是第二次世界大战期间把它用于原子弹的研制方面,具体地说,就是在电子计算机上对中子的行为进行随机抽样模拟,来推断所要求的参数。1946 年,物理学家冯·诺曼(Von Neumann)用随机抽样方法模拟了中子连锁反应,当时出于保密面将这种统计试验法以赌城蒙特卡罗命名为蒙特卡罗法。

蒙特卡罗法用于石油资源定量评价开始于本世纪 60 年代。美国于 1975 年完成的第二次全美石油资源评价的主要算法就是蒙特卡罗法。目前世界上各主要产油国及西方各大石油公司都把这种方法作为石油资源定量评价的重要方法之一,广泛应用于各含油气区的早、中期勘探阶段。我国应用该方法估算石油资源量开始于1979 年,现在各油田已普遍使用,并已成为以统计预测为主的应用软件评价系统的核心算法。

§2　随机数的产生和检验

应用蒙特卡罗法求数学、物理或工程技术问题的数值解时,要用数以千计、万计、甚至是百万计的随机数来解决不同分布的随机变量的抽样问题,因此,在求解计算之前必须先成功地解

决生成符合要求的随机数问题。或者说，在电子计算机上快速、经济地产生各种不同分布的随机数是蒙特卡罗法能够成功应用的基础。

所谓随机数就是随机变量 X 的抽样值 x，而抽样序列(有限值)$x_1, x_2, \cdots, x_n \cdots$叫做随机数序列。模拟石油资源量时，经常用到均匀分布的随机数，有时也用到正态分布的随机数。

早些年人们曾把事先造好的随机数整理成表的形式存入计算机中，使用时调用。也有人将放射性物质的随机放射源或物理噪声的随机噪声源与计算机联接，把随机的物理过程转变为随机数，这一种产生随机数的方法叫做物理随机数发生器法。上述的两种方法已不再使用。

目前，发展最快且使用最广泛的是用数学的方法产生随机数。但是，严格地讲，用数学方法不能产生真正的随机数，因此通常把用数学方法产生的随机数称作“伪随机数”。尽管如此，只要对伪随机数序列进行一系列严格的统计检验，证明其可以满足模拟计算的精度要求，则伪随机数就可以作为真正的随机数来使用。

为了满足模拟问题的实际需要，要求在计算机上产生随机数的速度要快，占用内存空间要少，产生的随机数序列要有足够长的周期，并且应当具有符合要求的概率统计性质。

从理论上讲，只要有一种连续分布的随机数，就可通过数学变换的方法产生其他分布的随机数。在连续分布函数中，〔0,1〕区间上标准均匀分布的随机变量是最简单、最基本的一种。因而有人把〔0,1〕上均匀分布的随机变量的抽样值称作随机数，而其他分布的随机数都可以借助均匀分布随机数来产生，所以说均匀分布随机数是随机抽样的基本工具。

一、随机数的产生

在计算机上产生随机数的数学方法有取中法，移位法和同余法等。前两种方法所产生的伪随机数序列的周期对初始值的依赖性很大，选的不好时，伪随机数序列的长度较短，不能满足模拟计算的需要。目前产生伪随机数序列比较好的方法是同余法中的乘同余法和混合同余法。

1. *乘同余法*

这种方法产生伪随机数序列的递推同余式为：

$$\begin{cases} x_{n+1} \equiv \alpha x_n \quad (\mathrm{mod} M) \\ r_{n+1} = x_{n+1}/M \end{cases} \tag{10-2}$$

式中 x_n、x_{n+1}—分别是第 n 次和第 $n+1$ 次产生的伪随机数；

α——乘子系数；

M——模；

r_n——〔0,1〕区间上的第 n 个伪随机数。

记号 $x_{n+1} \equiv \alpha x_n (\mathrm{mod} M)$叫以 M 为模的同余式，表示 x_{n+1}为 α 与 x_n 的积被 M 除后的余数部分。

乘同余法由 Lehmer 提出，他曾取 $M=10^8+1, \alpha=23$，初值 $x_0=47594118$，得到周期为 5882352 的 8 位十进制伪随机数序列，并对其中的 5000 个伪随机数进行了统计检验，其结果认为是满意的。

乘同余法的周期与初值 x_0 及 α 均有密切关系。设计算机字长为 k，$M=2^k$，若取 x_0 与 M 互素，α 与 M 符合一定的关系时，乘同余法的最大周期为 2^{k-2}。一般可取 $x_0=2a+1, \alpha=8q\pm3$，a、q 为正整数，最好取 $\alpha=5^{2l+1}$，其中 l 为使 $5^{2l+1}<2^k$ 成立的最大正整数。设 M 为 10 的幂次，则有：若 $M=10^k, k\geqslant5$，x_0 不是 2 或 5 的倍数，当且仅当 $\alpha(\mathrm{mod}200)$取下列值

3,11,13,17,21,27,29,37,53,59,61,67,69,77,83,91,109,117,123,131,133,139,141,147,163,171,173,179,181,187,189,197 之一时，乘同余法的周期为 $5\times10^{k-2}$。

2. 混合同余法

混合同余法产生伪随机数序列的递推同余式为：

$$\begin{cases} x_{n+1} \equiv \alpha x_n + \beta \quad (\mathrm{mod} M) \\ r_{n+1} = x_{n+1}/M \end{cases} \tag{10-3}$$

其中 x_n、x_{n+1}——分别为第 n 次、第 $n+1$ 次的伪随机数；

α——乘子系数；

β——增量；

M——模；

r_n——〔0,1〕区间上的第 n 个伪随机数。

【例 1】 当 $x_0=\alpha=\beta=6, M=10$ 时，由式(10-3)产生的伪随机数序列为 0.6,0.2,0.8,0.4,0.0,0.6,0.2,0.8,0.4,0.0,…。

【例 2】 设计算机字长为 k，当 $\alpha=1, M=2^k, x_0, \beta$ 任选时，式(10-3)产生的序列为 $x_{n+1}=x_n+\beta(\mathrm{mod}M)$，而〔0,1〕上的伪随机数 $r_{n+1}=x_{n+1}/M$。

由以上两例看出，序列的周期和统计性质与式(10-3)中参数 x_0、α、β 及 M 的选取有着密切的关系。例 1 中，伪随机数序列周期很短便产生循环，例 2 产生的序列虽有较长的周期，但它却不是随机的。适当地选取参数，可使式(10-3)产生的伪随机数序列周期长，而且统计性质符合需求。下面给出 α 和 β 的所有使得周期为 M 的选取方法，以便从中挑选符合统计性质的伪随机数序列。

一个混合同余伪随机数序列，达到周期为 M 的充要条件是：

① β 与 M 互素；

② 对每一个 M 的素因子 p，$\alpha-1$ 为 p 的倍数；

③ 若 M 是 4 的倍数，那么 $\alpha-1$ 是 4 的倍数。

在二进制的计算机上，若 $M=2^k$，那么应取 $\alpha=4q+1, \beta=2a+1$，x_0 为任意的非负整数，其中 q、a 为正整数。

二、伪随机数序列的统计检验

由前述可知，既便是目前产生伪随机数比较好的同余法，也不能在计算机上产生〔0,1〕区间上均匀分布的真随机数序列，只能是在参数选取恰当的条件下产生统计性质合乎要求，周期又有足够长度的伪随机数序列。这就是说，对于选定参数条件下所产生的一个随机数序列，只有经过严格的统计检验后，才能决定它是否可作〔0,1〕区间上均匀分布的随机数序列使用。一般说来，均匀性与独立性(随机性)是要对伪随机数序列进行统计检验的主要性质。当然，在解决不同类型的实际问题时，对各种统计性质要求不同，就要根据要求进行不同的检验。这里仅介绍检验伪随机数序列均匀性和独立性的几种检验方法。

1. 均匀性检验

对伪随机数序列的均匀性检验包括矩检验、频率检验和累积频率检验等。

(1) 矩检验

这是对伪随机数序列的各阶矩统计量的一种显著性检验。

若伪随机数序列的长度(序列内伪机数的个数)为 n，那么它的各阶矩为

$$m_k = \frac{1}{n}\sum_{i=1}^{n} r_i^k$$

式中 m_k——第 k 阶矩；

r_i^k——[0,1]区间上第 i 个伪随机数的 k 次方。

二阶中心矩(方差)为

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(r_i - m_1)^2 = m_2 - m_1^2$$

在理论上,标准均匀分布的各阶矩是已知的。k 阶矩 μ_k 和总体的方差 σ^2 分别为:

$$\mu_k = \frac{1}{k+1}; \quad \sigma^2 = \mu_2 - \mu_1^2 = \frac{1}{12}$$

假设生成的伪随机数符合均匀分布,则它的各阶矩 m_k、二阶中心矩 s^2 与标准均匀分布的各阶矩 μ_k、总体的方差 σ^2 应当一致。

若以 V 个伪随机数为一组,总共计算出 M 组一阶矩 m_{1j}、二阶矩 m_{2j}、二阶中心矩 s_j^2,并记它们的平均值为

$$\overline{m}_1 = \frac{1}{M}\sum_{j=1}^{M}m_{1j}; \quad \overline{m}_2 = \frac{1}{M}\sum_{j=1}^{M}m_{2j}; \quad \overline{s}^2 = \frac{1}{M}\sum_{j=1}^{M}s_j^2。$$

根据中心极限定理,当 $M\to\infty$ 时,$\overline{m}_2$,$\overline{m}_2$,$\overline{s}^2$ 的分布趋于标准正态分布,其平均值分别趋于 $\frac{1}{2}$、$\frac{1}{3}$、$\frac{1}{12}$,而总体方差应分别趋于 $\frac{1}{12V}$、$\frac{1}{45V}$、$\frac{1}{180V}$。

因而,可建立统计量

$$U_1 = (\overline{m}_1 - \frac{1}{2})/\sqrt{\frac{1}{12VM}}, U_2 = (\overline{m}_2 - \frac{1}{3})/\sqrt{\frac{4}{45VM}}, U_3 = (\overline{s}^2 - \frac{1}{12})/\sqrt{\frac{1}{180VM}}$$

U_1、U_2、U_3 近似服从正态分布,并且可以确定均匀性假设的临界区间 R,即

$$R_{\overline{m}_1}: \left[\frac{1}{2} - U_\alpha\sqrt{\frac{1}{12VM}}, \frac{1}{2} + U_\alpha\sqrt{\frac{1}{12VM}}\right]$$

$$R_{\overline{m}_2}: \left[\frac{1}{3} - U_\alpha\sqrt{\frac{4}{45VM}}, \frac{1}{3} + U_\alpha\sqrt{\frac{4}{45VM}}\right]$$

$$R_{\overline{s}^2}: \left[\frac{1}{12} - U_\alpha\sqrt{\frac{1}{180VM}}, \frac{1}{12} + U_\alpha\sqrt{\frac{1}{180VM}}\right]$$

当 $\overline{m}_1$、$\overline{m}_2$、$\overline{s}^2$ 大于对应的 R 的上界或小于 R 的下界时,则应否认均匀性假设。U_α 可从正态分布的双侧分位数表查得。

(2) 频率检验

这种检验又称拟合优度检验。如果伪随机数是均匀分布的,则可把[0,1]区间分成 k 个等子区间,一般取 $k=8,16,32$。此时可作假设 H_0:每个伪随机数属于第 i 个子区间的概率 $P_i=\frac{1}{k}$。也就是说,频率检验的目的在于检验每个子区间观测频数 n_i 与理论频数 $m_i=\frac{n}{k}$ 之间差别的显著性。为此,可建立 χ^2 检验统计量,即

$$\chi^2 = \frac{k}{n}\sum_{i=1}^{k}\left(n_i - \frac{n}{k}\right)^2 \tag{10-4}$$

式中 n——被检验的伪随机数个数;

k——[0,1]上的子区间数;

n_i——落入第 i 个子区间内的伪随机数个数。统计量 χ^2 的自由度为 $k-1$。

一般情况下,由于仅在一次试验中小概率事件是不容易发生的,如若发生,则认为实际观

测频数与理论频数之间相差显著，因而，可以否定假定 H_0，即认为伪随机数序列的分布是不均匀的。取检验置信水平 $\alpha=0.05$ 或 $\alpha=0.01$，若 $\chi^2 \geqslant \chi^2_{0.05}$ 称为差异显著，反之差异不显著；当 $\chi^2 \geqslant \chi^2_{0.01}$ 称为差异极显著，即伪随机数序列分布极不均匀。

(3) 累积频率检验

该检验亦称柯尔莫果罗夫拟合优度检验。设伪随机数序列的分布函数是 $F(r)$，而 $S_n(r)$ 是对该序列作 n 次独立观测获得的经验分布函数。根据柯尔莫果罗夫—斯米尔诺夫定理，对于任意 $\lambda>0$，则有等式

$$\lim_{n\to\infty} Q_n(\lambda) = \lim_{n\to\infty} p\left(D_n < \frac{\lambda}{\sqrt{n}}\right) = Q(\lambda) \tag{10-5}$$

其中

$$D_n = \sup_{0\leqslant r\leqslant 1} |F(r) - S_n(r)|, \; Q(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$$

因此，当试验次数 n 足够大时，就可以认为 $D_n<\frac{\lambda}{\sqrt{n}}$ 的概率 $P(D_n<\frac{\lambda}{\sqrt{n}})$ 趋于 $Q(\lambda)$。

如果 $D_n^{(0)}$ 是 n 次试验的 $|F(r)-S_n(r)|$ 的最大者，并且 $\lambda_0=\sqrt{n}D_n^{(0)}$，当

$$P(\sqrt{n}D_n \geqslant \lambda_0) = 1 - Q(\lambda_0) = \alpha$$

很小时，就发生了小概率事件，由此即可检验差异的显著性。这就是说，当 $\lambda_0 \geqslant \lambda_{0.05}$ 时，则可认为伪随机数序列的不均匀性是显著的，否则可认为是均匀分布的伪随机数序列。

2. 独立性检验

这种检验是对伪随机数的自相关性进行统计检验。它包括多种具体的检验方法，在此仅介绍其中的简单独立性检验和顺序检验。

(1) 简单独立性检验

简单独立性检验又称无重复列联检验。进行这种检验时，首先是把被检验的伪随机数序列等分为 ζ 与 η 两部分，并且要求任一伪随机数只能唯一地属于 ζ 或 η。

设 ζ、η 的取值分别为 ζ_t、$\eta_t(t=1,2,\cdots)$，把单位正方形划分为 k 行 h 列的网格，并把点 (ζ_t,η_t) 落在网格 (i,j) 内的频率记为 n_{ij}。若 ζ 与 η 相互独立，那么 ζ_t、η_t 同时出现的概率应为

$$p(\zeta_t,\eta_t) = p(\zeta_t)p(\eta_t) \tag{10-6}$$

因而，简单独立性检验就是以式(10-6)为假设 H_0，比较观测频数 n_{ij} 与理论频率 m_{ij} 之间的差异是否显著的一种检验。记

$$n_{i\cdot} = \sum_{j=1} n_{ij}, \quad n_{\cdot j} = \sum_{i=1} n_{ij}, \quad n = \sum_{i,j} n_{ij}$$

检验假设 $H_0: p(\zeta_t,\eta_t)=p(\zeta_t)p(\eta_t)$。给出检验水平 α，建立检验统计量

$$\chi^2 = \sum_{i=1}\sum_{j=1} \frac{(n_{ij}-m_{ij})^2}{m_{ij}} = \sum_{i=1}\sum_{j=1} \frac{n_{ij}^2}{m_{ij}} - n \tag{10-7}$$

因为

$$p(\zeta_t) \approx n_{i\cdot}/n, \quad p(\eta_t) \approx n_{\cdot j}/n$$

所以

$$m_{ij} = np(\zeta_t,\eta_t) \approx n_{i\cdot}n_{\cdot j}/n$$

又因

$$n_{i\cdot} = \sum_{j=1} n_{ij} = \sum_{j=1} m_{ij} \quad (i=1,2,\cdots,k)$$

$$n_{\cdot j} = \sum_{i=1} n_{ij} = \sum_{i=1} m_{ij} \quad (j = 1,2,\cdots,h)$$

所以自由度为

$$hk - (h + k - 1) = (h - 1)(k - 1) \tag{10-8}$$

由式(10-7)得到 χ^2 的值之后，便可与 χ_α^2 进行比较，若 $\chi^2 \geqslant \chi_{0.05}^2$，则称 ζ 与 η 之间显著相关，否则可认为 ζ 与 η 是独立的；若 $\chi^2 \geqslant x_{0.01}^2$，则称 ζ 与 η 之间极显著相关。

(2) 顺序检验

这种检验又称有重复列联检验。如果伪随机数序列是随机性的，就不会出现在某一类型的伪随机数之后总是出现另一种类型伪随机数的现象，如例 2 就是个非随机数性的伪随机数序列。

如果把序列中相应的两个随机数组成一个数字，便可构成二元频数表，而在所有网格内应当具有近似相等的频数。被检验的 n 个伪随机数可以组成 n 对伪随机数，并且 $n_{i\cdot} = n_{\cdot j}$。若 m_{ij} 是 n_{ij}的理论频数，则有

$$\chi^2 = \delta_2^2 - \delta_1^2 \tag{10-9}$$

其中

$$\delta_1^2 = \sum_{i=1}^{h} \frac{(n_{i\cdot} - m)^2}{m}, \quad m = \frac{n}{h}, \text{自由度为 } h - 1;$$

$$\delta_2^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(n_{ij} - m)^2}{m}, \quad m = \frac{n}{hk}, \text{自由度为 } hk - 1。$$

这里 h、k 的意义与前面的简单独立性检验相同。为了计算 δ_1^2、δ_2^2，需要把 n 个伪随机数按大小分为 N 种类型，即

$$\frac{j-1}{N} \leqslant r_i < \frac{j}{N} \quad (j = 1,2,\cdots,N)$$

$$\frac{k-1}{N} \leqslant r_{i+1} < \frac{k}{N} \quad (k = 1,2,\cdots,N)$$

分类后便可以计算出 δ_1^2、δ_2^2 以及 χ^2。若 $\chi^2 \geqslant \chi_{0.05}^2$，则称显著顺序相关，否则认为顺序不相关；若 $\chi^2 \geqslant \chi_{0.01}^2$，则称极显著顺序相关。

§3 随机变量的抽样

在随机数的基础上，对不同分布的随机变量进行抽样有多种方法。这里仅介绍经验分布函数抽样法、直接抽样法和变换抽样法。

一、经验分布函数抽样法

1. 经验分布函数

按照数学上的定义，随机变量 X 的取值不大于实数 x 的概率 $P(X \leqslant x)$ 为随机变量的分布函数，记为

$$F(x) = P(X \leqslant x)$$

所谓的经验分布函数是指直接由随机变量 X 的 n 个观测值 $x_1, x_2, \cdots, x_n$ 用统计的方法求得的分布函数，并将其记为 $F_n(x)$。在油气资源定量估算及其他地质研工作中，人们总是希望知道随机变量 X 的取值大于实数 x 的概率 $P(X > x)$，显然

$$P(X > x) = 1 - P(X \leqslant x) = 1 - F(x) \tag{10-10}$$

用 $F_n(x)$ 代替式(10-10)中的 $F(x)$，并记为

$$AF(x) = P(X > x) = 1 - F_n(x) \tag{10-11}$$

在此，我们称 $AF(x)$ 为石油资源评价中的经验分布函数，其曲线形态如图 10-1 所示。

2. 经验分布函数的构造方法

(1) 频率统计法

当随机变量 X 的观测值 $x_1, x_2, \cdots, x_n$ 为大样本($n \geqslant 30$)时，可以不受任何理论分布函数的制约，直接由随机变量的观测值用频率统计法求得它的经验分布函数。如果随机变量的观测值具有较好的代表性，那么这种由实际资料所得到的经验分布函数也就较好地描述出随机变量的统计特征。频率统计法求经验分布函数的步骤如下：

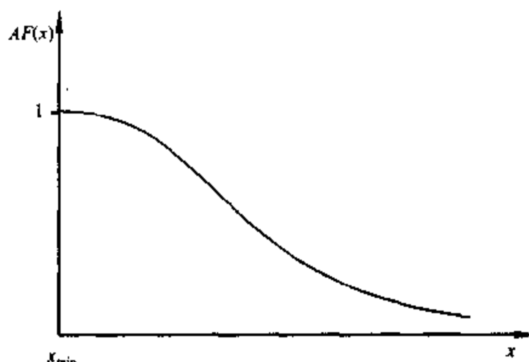


图 10-1 经验分布函数示意图

① 确定频率统计区间数

按平均落入每个统计区间内的观测值不少于 3～5 个的原则确定频率统计区间数 m。另外，为确保随机变量密度分布峰值的出现，m 最好为奇数。

② 计算区间间隔值

按式(10-12)求出 $m+1$ 区间间隔值 x_i

$$x_i = x_{\min} + \frac{x_{\max} - x_{\min}}{m}(i-1) \tag{10-12}$$

其中 $x_{\min}$ 和 $x_{\max}$ 分别是观测值的最小值和最大值。

③ 求经验分布函数

根据观测值 $x_1, x_2, \cdots, x_n$ 落入区间 (x_i, x_{i+1}) 内的频数 n_i 求得累积频率 f_i

$$f_i = \frac{1}{n}\sum_{j=i}^{m} n_j \qquad (i = 1, 2, \cdots, m) \tag{10-13}$$

得随机变量的经验分布函数

$$AF(x) = \begin{cases} 1, & x_1 \leqslant x \leqslant x_2 \\ f_2, & x_2 \leqslant x \leqslant x_3 \\ \vdots & \\ f_m, & x_{m-1} \leqslant x \leqslant x_m \\ 0, & x_m \leqslant x。 \end{cases}$$

在 x、$AF(x)$ 坐标系内，以 $(x_i + x_{i+1})/2$ 和 f_i 为坐标的分布函数曲线如图 10-1 所示。

(2) 等频率统计法

当随机变量的观测值为 10～30 个，而且又不知道随机变量的分布概型时，如果采用频率统计法，就会因统计区间个数少，使构造出的分布函数过于粗糙。在这种情况下，可认为每个观测值出现的概率是相等的。如果样本容量为 n，把观测值按大小顺序排成

$$x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$$

若 $x_k \leqslant x \leqslant x_{k+1}$，则不大于 x 的观测值的频率为 k/n。因而函数

$$F_n(x)=\begin{cases}0, & x<x_1\\ \dfrac{k}{n}, & x_k\leqslant x\leqslant x_{k+1}\\ 1, & x_n\leqslant x\end{cases}$$

经验分布函数为

$$AF(x)=1-F_n(x) \tag{10-14}$$

3. 经验分布函数的抽样

构造出经验分布函数 $AF(x)$后，利用随机数技术便可实现对随机变量的随机抽样了。

设 r_i 是随机数序列中的第 i 个随机数，那么随机抽样的过程是：首先在图 10-2 上确定点 $(x_{\min},r_i)$，通常称该点为随机入口点，而 r_i 称作随机入口值或概率入口值，然后过点 $(x_{\min},r_i)$作 x 轴的平行线交 $AF(x)$于点 (x_i,r_i)，交点的横坐标 x_i 即为随机变量的第 i 个观测值或第 i 个抽样值，常称它为随机抽样出口值。

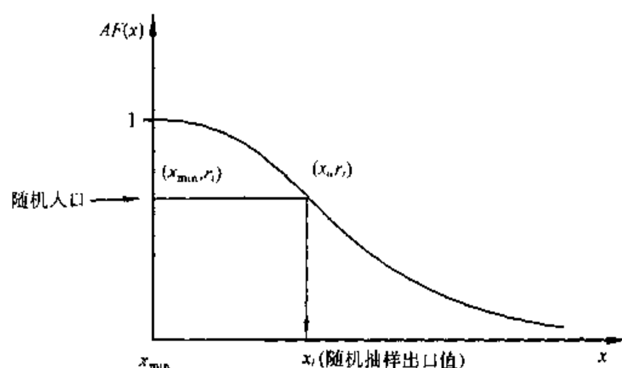


图 10-2　随机抽样过程示意图

二、直接抽样法

在理论上，[0,1]上均匀分布的随机变量与其他分布的随机变量有以下重要关系（简称随机变量理论关系）：

设随机变量 η 具有单调递增的连续分布函数 $F(x)$（或已给出分布密度 $f(x)$），则 $\zeta=F(\eta)$是[0,1]上均匀分布的随机变量。换句话说，如果 ζ 为[0,1]上均匀分布的随机变量，$F(x)$是某个随机变量的分布函数，且 $F(x)$为单调递增的连续函数，那么

$$\eta=F^{-1}(\zeta) \tag{10-15}$$

是以 $F(x)$为分布函数的随机变量。

由上述关系可知，只要随机变量具有连续单调递增的分布函数，$F^{-1}(x)$又能够用显式表示出来，就可以用均匀分布的随机抽样序列产生其他分布的随机抽样序列。但一般说来，式(10-15)中的 η 不能用 ζ 的显函数写出来。因此，直接抽样法仅适合对某一些分布概型的随机变量抽样。下面给出几个这种抽样方法抽样的例子。

【例 1】 由[0,1]上均匀分布的随机变量 ζ 产生$[a,b]$上均匀分布的随机变量 η。

$[a,b]$上均匀分布的随机变量 η 的密度函数为

$$f(x)=\begin{cases}\dfrac{1}{b-a} & (a\leqslant x\leqslant b)\\ 0, & \text{其他}\end{cases}$$

根据随机变量理论关系有

$$\zeta=\int_a^\eta \frac{1}{b-a}\mathrm{d}x=\frac{\eta-a}{b-a}$$

由此得

$$\eta=(b-a)\zeta+a \tag{10-16}$$

【例 2】 试由[0,1]上均匀分布的随机变量ζ产生服从三角分布的随机变量η。

当随机变量的密度函数为$f(x)=2x(0\leqslant x\leqslant 1)$时，称之为三角分布。根据随机变量理论关系有

$$\zeta=\int_0^\eta 2x\mathrm{d}x=\eta^2 \tag{10-17}$$

这表明[0,1]上均匀分布随机变量的平方根服从三角分布。

【例 3】 设ζ为[0,1]上均匀分布的随机变量，而服从指数分布的随机变量η的分布函数为

$$F(x)=1-e^{-\lambda x},\ (x>0,\lambda\text{为常数})$$

试把η用ζ表示出来。

由随机变量理论关系，可得

$$\zeta=1-e^{-\lambda\eta}$$

所以

$$\eta=-\frac{1}{\lambda}\ln(1-\zeta)$$

由于ζ是[0,1]上的均匀分布，而$1-\zeta$也是[0,1]上的均匀分布，故有

$$\eta=-\frac{1}{\lambda}\ln(\zeta)$$

【例 4】 标准正态分布的随机变量η的抽样序列不能用[0,1]上均匀分布随机变ζ的抽样序列表示出来。

根据随机变量理论关系，由标准正态分布的密度函数

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

虽然可得

$$\zeta=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\eta e^{-\frac{x^2}{2}}\mathrm{d}x$$

但是，η却不能用ζ的显函数形式表示出来。因此，无法由[0,1]上均匀分布随机变量的抽样序列产生标准正态分布随机变量的抽样序列。

三、变换抽样法

直接抽样法是对具有单调递增连续分布函数，并且又可用[0,1]区间上均匀分布随机变量ζ的显式直接表示出来的随机变量η的抽样来说的，而这里的变换抽样法则是针对存在单调递增连续分布函数，但它却不能表示为[0,1]上均匀分布随机变量ζ的显函数的随机变量η的抽样。

正态分布是常用的一类分布，它的抽样方法是统计试验法的重要内容。产生服从$N(0,1)$的抽样值ζ，又是获得服从$N(a,\sigma^2)$的抽样值η的基础。

设x_1,x_2是相互独立的[0,1]上的两个均匀分布随机数，作如下变换

$$
\begin{cases}
\zeta_1 = (-2\ln x_1)^{\frac{1}{2}}\cos 2\pi x_2 \\
\zeta_2 = (-2\ln x_1)^{\frac{1}{2}}\sin 2\pi x_2
\end{cases}
\tag{10-18}
$$

那么 ζ_1、ζ_2 则是两个相互独立的服从 $N(0,1)$分布的随机数。同样取 x_3、x_4，x_5、x_6，…又可产生 ζ_3、ζ_4，ζ_5、ζ_6，…。

依据中心极限定理，也可由[0,1]上均匀分布的随机数产生近似服从 $N(0,1)$分布的随机数。

设 $x_1,x_2,\cdots,x_n$ 是 n 个相互独立的[0,1]上均匀分布的随机数，那么 x_i 的期望和方差为

$$
E(x_i) = \frac{1}{2},\ D(x_i) = \frac{1}{12}
$$

根据中心极限定理，当 n 充分大时，

$$
\zeta_n = \left(\sum_{i=1}^{n} x_i - \frac{n}{2}\right) \Big/ \sqrt{\frac{n}{12}}
\tag{10-19}
$$

的分布渐近于 $N(0,1)$，故可把 ζ_n 近似看作服从标准正态分布的随机数。通常取 n 等于 8 或者 12，当 $n=12$ 时最为方便，此时

$$
\begin{aligned}
\zeta_{12} &= \sum_{i=1}^{12}(x_i - \frac{1}{2}) = \sum_{i=1}^{6}(x_{2i} - 1 + x_{2i-1}) \\
&= \sum_{i=1}^{6}(x_{2i} - x_{2i-1})。
\end{aligned}
\tag{10-20}
$$

如果随机变量 η 的数学期望为 $E(\eta)$，方差为 $D(\eta)>0$，那么随机变量

$$
\zeta = (\eta - E(\eta)) / \sqrt{D(\eta)}
$$

服从 $N(0,1)$分布，因此有

$$
\eta = \sqrt{D(\eta)} \cdot \zeta + E(\eta)
\tag{10-21}
$$

式(10-21)表明，在已知随机变量 η 的数学期望和方差的条件下，可由服从 $N(0,1)$分布的随机变量直接产生随机变量 η 的随机数。

§4 蒙特卡罗法预测含油区的石油资源总量

一、局部含油地质单元的石油资源量

1. 估算局部含油地质单元石油资源量的概率模型

局部含油地质单元是定量估算石油资源量的基本地质体，当采用不同的计算方法时，它的含义有所不同。例如，容积法计算石油资源量时的局部含油地质单元可以是一个油藏或一个油层；单储系数法计算石油资源量则以局部构造、断鼻、断块为局部含油地质单元；氯仿沥青法计算石油资源量时，局部含油地质单元可以是一个生油凹陷。但是，无论是采用哪一种计算方法，含油区中任何一个局部含油地质单元的石油资源量常用计算公式，都可归结为一些地质参数与经验系数的连乘。

如果一个含油区中总共有 m 个局部含油地质单元，那么其中第 j 个局部含油地质单元的石油资源量

$$
Q_j = \prod_{k=1}^{l} C_{jk} \prod_{i=1}^{n} X_{ji} = k_j \prod_{i=1}^{n} X_{ji}
\tag{10-22}
$$

其中　k_j——第 j 个局部含油地质单元中 l 个地质常数与经验系数的积；

X_{ji}——第 j 个局部含油地质单元中与石油资源量有关的第 i 个地质参数(随机变量)。

式(10-22)是估算第 j 个局部含油地质单元石油资源量的概率模型。

2. *石油资源量的计算*

为了统计第 j 个局部含油地质单元石油资源量的累积频率,在抽样计算之前,首先要由实测值求出第 j 个局部含油地质单元石油资源量的最大可能值 $q_{j\max}$、最小可能值 $q_{j\min}$ 及累积频率区间间隔值 q_{jh},即

$$\begin{cases} q_{j\max} = k_j \prod_{i=1}^{n} x_{ji\max} \\ q_{j\min} = k_j \prod_{i=1}^{n} x_{ji\min} \end{cases} \quad (j = 1,2,\cdots,m) \tag{10-23}$$

$$q_{jh} = q_{j\min} + \frac{q_{j\max} - q_{j\min}}{v}(h-1) \tag{10-24}$$

$$(j = 1,2,\cdots,m; h = 1,2,\cdots,v+1)$$

式(10-23)及式(10-24)中 $x_{ji\max}$ 及 $x_{ji\min}$ 分别是第 j 个局部含油地质单元中第 i 个地质变量的最大、最小观测值,而 q_{jh} 是累积频率的第 h 个区间间隔值,v 是区间数。

选择§3中的经验分布函数抽样法,直接抽样法或者变换抽样法,获得局部含油地质单元中 n 个随机变量 $X_{j1}, X_{j2}, \cdots, X_{jn}$ 的第 $g(g=1,2,\cdots,N)$ 次抽样值 $x_{j1g}, \cdots, x_{jng}$,并将其代人石油资源量概率模型式(10-22),计算出第 j 个局部含油地质单元石油资源量的估计值 $q_{jg}(g=1,2,\cdots,N)$,最后用频率统计法对局部含油地质单元石油资源的 N 个估计值进行整理,就可以求出含油区中第 j 个局部含油地质单元石油资源量 Q_j 的分布函数 $AF(q_j)$。

在此需要指出的是:如果是采用经验分布函数抽样法,那么计算机内存储的并非是经验分布函数的整条分布曲线,而是曲线上的 U 个型值点 $(x_{jip}, AF(x_{jip}))(p=1,2,\cdots,u)$。当抽样落入型值点 $(x_{jip}, AF(x_{jip}))$ 与 $(x_{jip+1}, AF(_{jip-1}))$ 之间时,如图10-3所示,随机变量 X_{ji} 在抽样点的

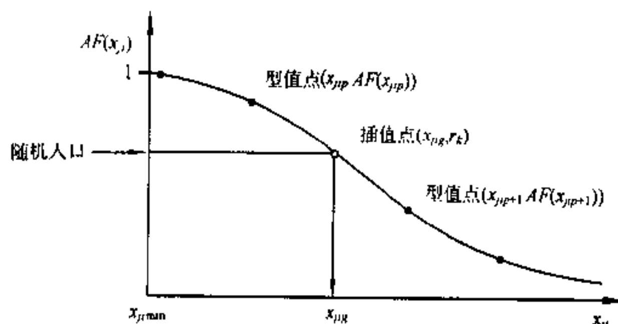


图10-3 线性插值示意图

值用插值(线性或非线性)法求出。若按线性插值方法计算,那么插值计算公式为:

$$x_{jig} = x_{jip} + \frac{(x_{jip+1} - x_{jip})(r_k - AF(x_{jip}))}{AF(x_{jip+1}) - AF(x_{jip})} \tag{10-25}$$

$$(j = 1,2,\cdots,m; i = 1,2,\cdots,n; p = 1,2,\cdots,U; g = 1,2,\cdots,N)$$

上式中 r_k 是随机数序列中的第 k 个随机数。

二、含油区的石油资源总量

1. 含油区石油资源量的概率模型

这里所说的含油区可以是一个地质凹陷，一个地质坳陷、一个沉积盆地、以至超越盆地的一个范围较大的含油地区。因此，含油区的石油资源总量可能需要多级累加才能得到。例如，含油区是一个沉积盆地，局部含油地质单元是地质圈闭时，则要根据地质圈闭的石油资源量先求出地质凹陷的石油资源量，再由地质凹陷的石油资源量求出地质坳陷的石油资源量，最后由地质坳陷的石油资源量得到全盆地的石油资源总量，即

$$Q = \sum_{j_1=1}^{m_1}\sum_{j_2=1}^{m_2}\cdots\sum_{j_k=1}^{m_k} Q_{j_1j_2\cdots j_k} \tag{10-26}$$

上式为含油区石油资源总量的一般概率模型。

式中　Q——含油区石油资源总量；

$Q_{j_1j_2\cdots j_k}$——第 $j_1,j_2,\cdots,j_k$ 级分区中的局部含油地质单元的石油资源量。

2. 含油区的石油资源总量

为了化简问题，这里仅以局部含油地质单元一次求和作为含油区的石油资源总量。因此，式(10-26)化简为

$$Q = \sum_{j=1}^{m} Q_j \tag{10-27}$$

为了统计含油区石油资源总量的累积频率，在对总资源量抽样计算之前，如同计算第 j 个局部含油地质单元的石油资源量 Q_j 一样，也要先求出含油区石油资源总量 Q 出现的最大可能值、最小可能值及总资源累积频率区间间隔值，即：

$$q_{max} = \sum_{j=1}^{m} q_{jmax},\ q_{min} = \sum_{j=1}^{m} q_{jmin} \tag{10-28}$$

及

$$q_l = q_{min} + ((q_{max} - q_{min})/v)(l-1) \tag{10-29}$$

式(10-29)中，q_{max}、q_{min}是含油区内石油资源总量的最大、最小可能值；q_l 是累积频率的第 l 个区间间隔值；v 为累积频率区间数。

对含油区中 m 个局部含油地质单元的石油资源量分布函数 $AF(q_j)$进行随机抽样，并把第 $i(i=1,2,\cdots,N)$次的抽样值 $q_{1i},q_{2i},\cdots,q_{mi}$代入式(10-27)，得到含油区石油资源总量 Q 的 N 个估计值 $q_1,q_2,\cdots,q_N$。最后再应用频率统计法由 $q_1,q_2,\cdots,q_N$ 求出含油区石油资源总量的分布函数 $AF(q)$。

三、石油资源总量分布函数的正态化及内插整理

1. 石油资源总量分布函数的正态化

由若干个局部含油地质单元的石油资源量分布函数经过多次加法运算，求得的含油区石油资源总量分布函数 $AF(q)$的形态趋向正态分布。产生这种现象的原因是由 m 个抽样值累加和的均匀化所致，它服从中心极限定理。

根据李雅普诺夫(Liapunov)定理有：如果随机变量 $Q_1,Q_2,\cdots,Q_n,\cdots$ 相互独立，它们具有有限的数学期望 $E(Q_j)$和方差 $D(Q_j)$，即

$$E(Q_j) = a_j, D(Q_j) = \sigma_j^2 \neq 0 \qquad (j=1,2,\cdots,n,\cdots)$$

记：

$$B_n^2 = \sum_{j=1}^{n} \sigma_j^2$$

若存在正整数 δ，使得当 $n \to \infty$ 时

$$\frac{1}{B_n^{2+\delta}} \sum_{j=1}^{n} E|Q_j - a_j|^{2+\delta} \to 0$$

则随机变量

$$Z_n = \left(\sum_{j=1}^{n} Q_j - \sum_{j=1}^{n} a_j \right) / B_n$$

的分布函数 $F_n(q)$ 对于任意 q，满足：

$$\lim_{n\to\infty} F_n(q) = \lim_{n\to\infty} p(Z_n \leqslant q) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q} e^{-\frac{t^2}{2}} \mathrm{d}t$$

亦即，当 n 充分大时，随机变量 Z_n 将服从正态分布 $N(0,1)$。

由此得出，当 n 充分大时

$$\sum_{j=1}^{n} Q_j = B_n Z_n + \sum_{j=1}^{n} a_j$$

将服从正态分布 $N\left(\sum_{j=1}^{n} a_j, \sum_{j=1}^{n} \sigma_j^2 \right)$。

上述结论指出，无论各个局部含油单元石油资源量 Q_j 的分布函数 $AF(q_j)$ 具有怎样的分布，只要满足上述定理的条件，则当累加次数 n 充分大时，含油区石油资源总量 $\sum_{j=1}^{m} Q_j$ 的分布函数 $AF(q)$ 将会近似地服从正态分布。

2. 石油资源总量分布函数的内插整理

经过多次累加后的含油区石油资源总量的分布函数，其资源量的实际变化范围要缩小，如图 10-4 所示。现以 $m=2$ 为例说明如下：

假如 Q_1 及 Q_2 的分布函数 $AF(q_1)$ 及 $AF(q_2)$ 都属于正态分布，经过抽样值的加法运算后得到 Q 的分布函数 $AF(q)$。此时，只在概率为 50% 处 $q=q_1+q_2$；当概率 >50% 时，$q>q_1+q_2$，而概率 <50% 时，$q<q_1+q_2$。

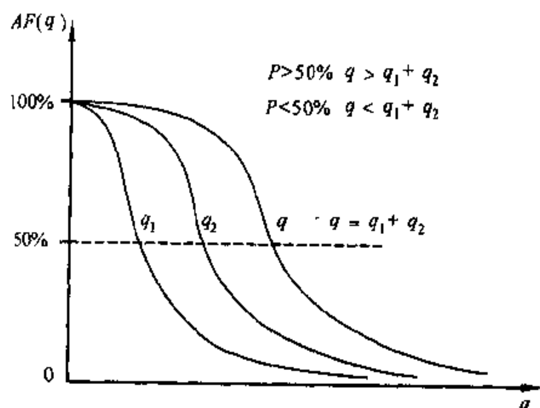


图 10-4 石油资源总量区间范围收缩现象

上述现象的出现与否以及发生的程度与被累加的分布函数 $AF(q_j)$ 的概型及累加的次数有关。如果被累加的分布函数均为正态分布，那么区间两侧的收缩程度相等，累加次数越多，区间收缩得也就越明显，因此，在石油资源量分布区间的大值一侧，含油区的石油资源总量就小于 m 个局部含油地质单元石油资源量的累加和，而在石油资源量分布区间的小值一侧，含油区的石油资源总量就大于 m 个局部含油地质单元石油资源量的累加和。

由此，经过抽样模拟计算后得到的含油区石油资源总量，其分布函数 $AF(q)$ 的左端，往往会出现很多个概率为 1 的数值点；而在右端，会出现很多个概率为 0 的数值点。因为分布函数 $AF(q)$ 的两端只应保留概率为 1 及 0 的点各一个，为此可通过区间 $[q_{\max}, q_{\min}]$ 内部插值计算，

去掉多余的概率为1及0的数值点，求出有效区间$[q'_{max}, q'_{min}]$范围内对应的分布函数$AF'(q)$。但是，这种内插整理计算，需要在全部累加计算完成后一次进行，以防止多次内插计算时产生误差的累积传播。

四、地质风险分析

由于石油勘探的未来成效具有不确定性，因而特别需要对估算的石油资源量进行风险分析。所谓风险就是失败的机会。石油勘探中的风险是多种多样的。如勘探地区是否具备形成油气藏地质条件的地质风险；在已具备形成油气藏地质条件的含油区内，经过勘探能否找到一定规模油气藏的勘探风险；勘探后已经发现的油气藏是否具备开采价值的经济风险；石油勘探过程中人与设备是否安全的环境风险；对于勘探地区，特别是海域大陆架地区是否有国际争议的政治风险等等。

显而易见，上述种种风险都会对石油勘探起着决定性影响，因而风险分析是石油资源评价工作中不可缺少的重要环节。

石油地质勘探的专业人员必须认真作好地质风险分析。在实际工作中，地质风险分析可以在不同的层次进行，例如，单一地质圈闭的风险分析，一组地质圈闭（国外对地质条件相似的一组地质圈闭称作一个勘探层）的风险分析，一个油气聚集带的风险分析、以及整个含油气盆地的风险分析等等。

一般情况下，地质风险分析大都从地质圈闭算起，其计算公式为

$$K = 1 - \prod_{i=1}^{n}(1 - k_i) \tag{10-30}$$

式中　K——风险值；

k_i——第i个地质因素的风险值。

例如，地质勘探人员用如下的容积法公式估算一个地质圈闭的石油储量

$$Q = S \cdot H \cdot \Phi \cdot D \cdot W \tag{10-31}$$

式中　S——含油面积；

H——储集层厚度；

Φ——储集层孔隙度；

D——石油充满系数；

W——采收率。

风险分析时，要由熟悉探区情况的地质人员对上述5个地质参数逐个进行分析论证。就一般的地质概念而论，上述5个地质参数中，含油面积S的风险常常决定于地质调查或地震勘探资料的可靠性；储集层厚度H的风险受岩性岩相变化的影响；储集层孔隙度Φ的风险决定于储集层孔隙是否有次生改造或后期充填的影响；石油充满系数D的风险可能受生油岩的成熟程度及油气运移通道的制约；而采收率W的风险则与原油性质及驱动类型有关。

经过认真分析论证后，要对5个地质参数给定风险值k，一般用小数表示。而$p=(1-k)$可称作保险值。给定风险值在目前还没有一套完善的方法，一种方法是由地质人员凭经验人为确定；另一种方法是根据含油气地质条件相似的邻区资料，通过统计分析确定。

例如，某个地质圈闭经过分析后给出如下风险值，见表10-1。

表 10-1 单一地质圈闭的风险数据表

地质参数	风险值/k	保险值/$(1-k)$
含油面积(S)	0.0	1.0
储集层厚度(H)	0.5	0.5
储集层孔隙度(Φ)	0.0	1.0
石油充满系数(D)	0.3	0.7
采收率(W)	0.0	1.0

按(10-30)式计算便可以得到这个地质圈闭的风险系数 K,即

$$\begin{aligned}K &= 1-\prod_{i=1}^{5}(1-k_i)\\ &= 1-(1.0\times 0.5\times 1.0\times 0.7\times 1.0)\\ &= 1-0.35\\ &= 0.65\end{aligned}$$

而保险值为 0.35。

如果这个地质圈闭的石油储量分布函数为 $AF(q_j)$,经过风险分析后,分布函数的每个值都要下降到原来概率 35%的地方,见图 10-5。

图 10-5 中的曲线①为风险分析前的石油储量分布函数;曲线②为风险分析后的石油储量分布函数。

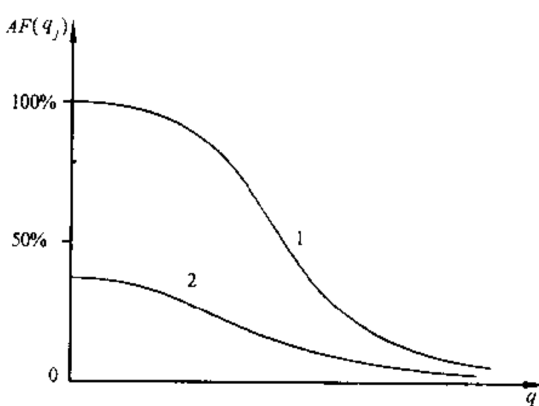


图 10-5 风险分析后的石油储量分布函数

如果一组在含油气地质条件上相类似的可能含油地质圈闭,总共 10 个。这组地质圈闭如按(10-31)式估算石油储量时,经过地质人员认真分析后,其各项地质参数的风险值见表 10-2。

试问在这一组地质圈闭中,至少获得一个油藏的可能性有多大?这里有两种计算方法。

表 10-2 一组可能含油圈闭的风险数据表

圈闭序号 / $(1-k)$ / 地质参数	1	2	3	4	5	6	7	8	9	10
S	1.0	1.0	0.5	1.0	0.5	1.0	1.0	1.0	1.0	1.0
H	0.5	1.0	1.0	1.0	1.0	0.5	1.0	0.5	1.0	1.0
Φ	1.0	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5
D	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
W	1.0	1.0	1.0	0.5	1.0	1.0	0.5	1.0	0.5	1.0

(1) 单独计算每个地质圈闭的风险系数,再计算一组地质圈闭的风险系数 K

$$K=\prod_{j=1}^{10}\left[1-\prod_{j=1}^{5}(1-k_{ji})\right]=(1-0.25)^{10}=0.0563$$

最后再求出在这组地质圈闭中发现一个油藏的保险系数 p

$$p=1-k=1-0.0563=0.9437$$

即，发现一个油藏的可能性为94.37%。

(2) 以最不利的地质因素，即以充满系数 D 的风险值来计算发现一个油藏的风险系数 K

$$K=\prod_{i=1}^{10}k_i=(1-0.5)^{10}=0.00098$$

最后再求出在这组地质圈闭中发现一个油藏的保险系数 p

$$p=(1-0.00098)\times 0.5=0.4951$$

即发现一个油藏的可能性为49.51%。

通过计算，可见这两种计算方法给出的结果并不一样。为什么用第二种算法得到的保险系数偏小呢？其原因是在这5个地质因素中，每个圈闭的充满系数都存在风险，可见对于这组圈闭来说，充满系数的风险最大，所以其保险系数小。

至于什么情况下用第一种算法，什么情况下用第二种算法，要根据探区的地质情况由地质人员选定。

风险分析时，如果地质参数间具有多层次结构，则要计算复合地质风险值。例如，某探区的地质风险决定于生油条件和储油条件。而生油条件与生油层厚度及生油相带有关；储油条件与储集层厚度及储集层相带有关，见表10-3。

表10-3 复合地质风险数据表

基础地质因素	风险值/k_{ij}	保险值/$(1-k_{ji})$	组合地质因素	风险值/k_j	保险值/$(1-k_j)$
生油层厚度	0.4	0.6	生油条件	0.4	0.6
生油层相带	0.0	1.0			
储集层厚度	0.1	0.9	储油条伯	0.37	0.63
储集层相带	0.3	0.7			

复合地质风险可按下式计算：

$$\begin{aligned}K&=1-\prod_{j=1}^{m}\left\{1-\left[1-\prod_{i=1}^{n}(1-k_{ji})\right]\right\}\\&=1-\prod_{j=1}^{m}\prod_{i=1}^{n}(1-k_{ji})\end{aligned}\qquad(10\text{-}32)$$

式中 K——复合风险值；

k_{ji}——第 j 项组合地质因素的第 i 个基础地质因素的风险值。

表10-3中的地质风险数据，按式(10-32)计算，其复合地质风险值如下：

$$\begin{aligned}K&=1-(0.6\times 0.1)\times(0.9\times 0.7)\\&=1-0.378\\&=0.622\end{aligned}$$

如果地质数据间的结构层次不止两层，则复合地质风险值可按下式计算：

$$K=1-\prod_{j_1=1}^{m_1}\prod_{j_2=1}^{m_2}\cdots\prod_{i=1}^{n}(1-k_{j_1j_2\cdots i})\qquad(10\text{-}33)$$

从以上计算中可以看出，求逆概率是风险分析的主要算法。

五、风险分析后的石油资源量求和计算

当每个局部含油地质单元的石油资源量都经过风险分析后，求含油区的石油资源总量时，做法是：用随机数 r，对第 j 个局部含油地质单元进行抽样计算，若随机数的值大于保险系数 $(1-k)$，由于入口值不能与分布函数 $AF(q_j)$ 相交，该次对第 j 个局部含油地质单元的抽样结果

q_{jw}应等于 0；只有在随机数的值小于或等于保险系数$(1-k)$时，才能通过插值计算得到抽样结果 q_{jw}，见图 10-6。

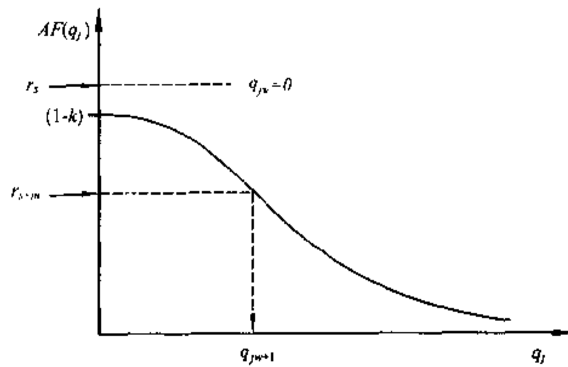


图 10-6　风险分析后的抽样计算

从图 10-6 可以看出，用[0,1]区间上的第 s 个随机数 r_s 对第 j 个局部含油地质单元抽样时，若 $r_s>(1-k)$，应令 $q_{jw}=0$。由于含油区共有 m 个局部含油地质单元，下次再对第 j 个局部含油地质单元进行抽样时，所用的随机数应为序列中的第 $s+m$ 个，若 $r_{s+m}<(1-k)$，则可由插值计算得到 q_{jw+1}。

图 10-7 是风险分析后，由局部含油地质单元求含油区石油资源总量的示意图。

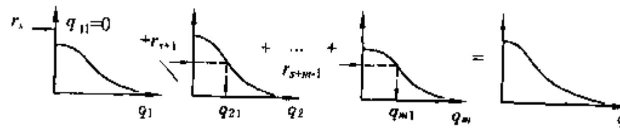


图 10-7　风险分析后计算石油资源总量的示意图

由图 10-7 可知，假如第 s 个随机数 r_s 大于第一个局部含油地质单元的保险系数$(1-k_1)$，则第一次抽样值 q_{11}应等于 0；而第 $s+1$ 个随机数 r_{s+1}小于第二个局部含油地质单元的保险系数$(1-k_2)$，则第二个局部含油地质单元的第一次抽样值 q_{21}不等于 0；……；直到用第 $s+m-1$ 个随机数 r_{s+m-1}对第 m 个局部含油地质单元进行第一次抽样，而得到 q_{m1}。

将这 m 个抽样值累加，则可得到含油区石油资源总量的一个随机估计值 q_w，即

$$q_w = \sum_{j=1}^{m} q_{jw}$$

请注意，上式中的 q_{jw}有时为 0。

如果一共抽样计算 g 次，则可得到含油区石油资源总量的 g 个随机估计值。最后用频率统计法可以求出含油区石油资源总量 Q 的分布函数 $AF(q)$。

由风险分析后的局部含油地质单元石油资源量的分布函数 $AF(q_j)$求得的含油区石油资源总量的分布函数 $AF(q)$，其曲线形态多呈偏态分布。出现这种现象的原因前面已多次提到，是由于风险分析后，当随机数的值大于$(1-k)$时，使得局部含油地质单元的一些抽样值为 0 之故。同时也造成许多抽样和的值偏小，因而使含油区石油资源总量区间小值一侧抽样和的频率增大，而向大值一侧抽样和的频率迅速变小。所以分布函数曲线的高峰偏向小值一侧、而大值一侧曲线缓慢下降，即呈现偏态的长尾分布，如图 10-7 所示。

当然，只是在被累加的局部含油地质单元的数量不太多的情况下，才会出现偏态的长尾分布。当被累加的局部含油地质单元的数量充分大时，含油区石油资源总量的分布函数必将按中心极限定理趋向正态分布。

§5　蒙特卡罗法 FORTRAN 源程序

一个蒙特卡罗法估算石油资源量的程序，应具备全面的计算和输出功能。它应是一个功能很强的软件包。在程序设计时，一般应考虑如下五个方面：

(1) 应当适用于任何形式的石油资源量计算公式。计算公式中可以包括任意个随机变量及经验系数。

(2) 应当具备在原始数据的数量不同的条件下，构造随机变量分布函数的功能。

(3) 应当能够估算任意个局部含油地质单元石油资源的合计资源量。而合计资源量又可分为若干个级别。例如有圈闭、构造带、坳陷、盆地石油资源总量，则分为 4 个级别。因而程序不仅能估算单个局部含油地质单元的石油资源量，而且要具备多级求和的功能。

(4) 应具备对石油资源量进行各种风险分析的功能。

(5) 能够输出随机变量的分布函数曲线以及各级石油资源量的分布曲线和计算结果，特别是要输出石油资源量的汇总表。

鉴于上述，一个完整的估算石油资源量的计算程序就相当长，考虑到教学上的适用，这里给出的程序是估算石油资源量的蒙特卡罗法的基础性程序。但该程序适用于任何形式的石油资源量计算公式。当原始数据较多时，可以用频率统计法求随机变量的分布函数，当数据较少但已知服从正态分布时，可由变换抽样法求得随机变量的分布函数。输出随机变量分布函数曲线、概率密度曲线以及局部含油地质单元和含油区石油资源总量分布函数曲线。

一、主要符号说明

k——含油区中局部含油地质单元总数；

l——石油资源量计算公式中的随机变量个数；

l1——石油资源量计算公式中经验系数与常数的个数；

n——随机变量的原始数据个数；

m——随机变量分布函数的区间数；

m1——随机变量分布函数区间间隔值个数；

kk——资源量分布函数横坐标控制参数。当 kk=1 时为乘积区间，kk=0 时为自然对数区间；

md——混合同余法产生随机数的模；

la——乘子系数；

ng——混合同余法产生随机数的初值；

x(102)——局部油地质单元的石油资源量；

xx(200)——随机变量原始数据；

af(101)——存放局部含油地质单元石油资源量概率的一维数组；

sx(102)——存放含油区石油资源总量的数组；

saf(101)——存放含油区石油资源总量概率的数组；

nx(101)——频数计数单元；

p(101)——存放频率的数组；

dxI(10)——每个随机变量取值范围的百分之一；

xi(10,13)——二维数组；行为随机变量号，列为随机变量的区间间隔值；

afi(10,12)——二维数组，行为随机变量号，列为随机变量的区间累积频率；

mpy——求随机变量分布函数并估算局部含油地质单元石油资源量子程序名；

add——估算含油区石油资源总量子程序名；

map——绘制分布函数及概率密度曲线子程序名。

二、程序使用说明

1. 数据文件

运行本程序之前，先把估算局部含油地质单元石油资源量的 n 个随机变量的原始数据及经验系数和常数建立n+1数据文件。若含油区有 m 个局部含油地质单元，那么一共要建立m(n+1)个数据文件。

2. 程序运行

程序运行时，由键盘输入绘图坐标原点(x0,y0)、资源量分布函数横坐标控制参数 kk、抽样次数 km、局部含油地质单元数 k、随机变量个数 l、经验系数及常数的个数 l1、绘图文件名，然后确定建立随机变量分布函数的方法，并依次输入局部含油地质单元随机变量的原始数据文件名。

3. 主要输出结果

程序运行结束，输出局部含油地质单元随机变量的分布函数及概率密度曲线、局部含油地质单元石油资源量和含油区石油资源总量的分布函数曲线，以及不同概率下的石油资源量。

三、源程序

```
$debug
        dimension x(102),af(101),sx(102),saf(101)
        common xx(200),xi(10,13),afi(10,12),nx(101),
     #  p(101),w(102),r(101),k1(10),x1(10),dxi(10)
        common/xy/x0,y0,xl,ny(2),bt(2)
        data md,la,xl/524288,3125,10./
        data ny(1),ny(2),bt(1),bt(2),nyy,nyk/11,13,7,3,11,23/
        write(*,*)'input x0,y0'
        read(*,*) x0,y0
        write(*,*)'input kk,km'
        read(*,*) kk,km
        write(*,*)'input k,L,L1'
        read(*,*) k,L,l1
        write(*,100) k,l,l1,kk,km
100     format(//60('*')//4x,'k=',i3,5x,'l=',i3,5x,'l1=',i3,
     #  5x,'kk=',i3,5x,'km=',i8)
        call in
        call fact(10.)
        do 130 i=1,k
```

```
      write( * ,110) i
110   format(//60('*')//4x,11('*')/4x,
   #  5h * i= ,i3,3h * /4x,'*',9('*'),'*')
      if(i.eq.1) go to 120
      call mpy(x,af,l,l1,md,la,nyy,kk,km)
      call add(x,af,sx,saf,i,k,md,la,nyk,km)
      go to 130
120   call mpy(sx,saf,l,l1,md,la,nyy,kk,km)
130   continue
      stop
      end

      subroutine mpy(x,af,l,l1,md,la,nyy,kk,km)
      dimension x(102),af(101),pp(101),sj(2),f(101)
      common xx(200),xi(10,13),afi(10,12),nx(101),
   #  p(101),w(102),r(101),k1(10),x1(10),dxi(10)
      common/xy/x0,y0,xl,ny(2),bt(2)
      character * 10 type
      write( * , * ) 'input type '
      read( * ,'(a)') type
      do 300 k=1,L
      write( * , * ) ' input n,nr '
      read( * , * ) n,nr
      open(nr,file=' ')
      read(nr, * ) (xx(i),i=1,n)
      write( * ,100) k,n,(xx(i),i=1,n)
100   format(//5x,3hxx(,i2,1h ,i4,1h)/(5x,10f6.2))
      m=n/5
      Ln=0
      if(type.eq.'y'.or.type.eq.'Y') m=km/5
      if(mod(m,2).eq.0) m=m-1
      if(m.gt.11) m=11
      if(m.lt.5) m=5
      m1=m+1
      m2=m+2
      if(m.eq.5) mm=1
      if(m.eq.7) mm=2
      if(m.eq.9) mm=3
      if(m.eq.11) mm=4
      xmax=xx(1)
```

```
      xmin=xx(1)
      do 105 i=1,n
      if(xx(i).gt.xmax) xmax=xx(i)
105   if(xx(i).le.xmin) xmin=xx(i)
      dx=(xmax-xmin)/float(m)
      dxi(k)=(xmax-xmin)/100.
      do 110 i=1,m2
110   xi(k,i)=xmin+dx*(i-1)
      xi(k,1)=xi(k,1)-dx/2
      xi(k,m1)=xi(k,m1)+dx/2
      do 120 i=1,m1
120   nx(i)=0
      if(type.eq.'y'.or.type.eq.'Y') then
      ex=0.
      sgm=0.
      do 130 i=1,n
      ex=ex+xx(i)
130   sgm=sgm+xx(i)**2
      ex=ex/n
      sgm=sqrt((sgm-n*ex**2)/(n-1))
      do 170 i=1,km
      do 140 j=1,2
      ly=la*ny(j)+bt(j)
      ny(j)=mod(ly,md)
140   sj(j)=float(ny(j))/float(md)
      st=sqrt(-2.*alog(sj(1)))
      r1=st*cos(2*3.1416*sj(2))
      r2=st*sin(2*3.1416*sj(2))
      r1=r1*sgm+ex
      r2=r2*sgm+ex
      do 160 j=1,m1
      if(r1.gt.xi(k,j).and.r1.le.xi(k,j+1)) then
      nx(j)=nx(j)+1
      Ln=ln+1
      end if
      if(r2.gt.xi(k,j).and.r2.le.xi(k,j-1)) then
      nx(j)=nx(j)+1
      Ln=ln+1
      end if
160   continue
```

```
170       continue
          s=0.
          do 180 i=1,m1
          p(i)=float(nx(i))/float(Ln)
          afi(k,i)=1.0-s
180       s=s+p(i)
          else
          s=0.0
          do 200 i=1,m1
          do 190 j=1,n
          if(xx(j).gt.xi(k,i).and.xx(j).le.xi(k,i+1)) then
          nx(i)=nx(i)+1
          end if
190       continue
          p(i)=float(nx(i))/float(n)
          afi(k,i)=1.0-s
          s=s+p(i)
200       continue
          end if
          write(*,210)k,m1,(p(i),i=1,m1)
210       format(/5x,'p(',i2,',',i2,')'/4x,12f6.4)
          write(*,220) k,m1,(afi(k,i),i=1,m1)
220       format(/5x,'afi(',i2,',',i2,')'/4x,12f6.4)
          xi(k,1)=xi(k,1)+dx/2.
          xi(k,m1)=xi(k,m1)-dx/2.
          do 230 i=1,m1
          x(i)=xi(k,i)
          f(i)=2*afi(k,i)
230       pp(i)=2*p(i)
          x0=1.
          call map(x,f,m1,1)
          x0=x0+xl+3.
          call map(x,pp,m1,1)
          y0=y0+3.*f(1)
          k1(k)=m1
          write(*,240) k,m1
240       format(5x,'xi(',i2,',',i2,')')
          if(mm.eq.1) write(*,250)
     #    (xi(k,i),i=1,m1,2),(xi(k,i),i=2,m1,2)
250       format(1x,3f8.2/4x,3f8.2)
```

```
        if(mm.eq.2) write(*,260)
     #  (xi(k,i),i=1,m1,2),(xi(k,i),i=2,m1,2)
260     format(1x,4f8.2/4x,4f8.2)
        if(mm.eq.3) write(*,270)
     #  (xi(k,i),i=1,m1,2),(xi(k,i),i=2,m1,2)
270     format(1x,5f8.2/4x,5f8.2)
        if(mm.eq.4) write(*,280)
     #  (xi(k,i),i=1,m1,2),(xi(k,i),i=2,m1,2)
280     format(1x,6f8.2/4x,6f8.2)
300     continue
        write(*,*) 'input nnr'
        read(*,*) nnr
        open(nnr,file=' ')
        read(nnr,*) (x1(j),j=1,L1)
        write(*,310) l1,(x1(j),j=1,l1)
310     format(//5x,'x1(',i2,')'/(5x,f10.4))
        s=1.0
        do 320 i=1,l1
320     s=s*x1(i)
        if(kk.eq.0) go to 350
        do 340 i=1,102
        x(i)=1.0
        do 330 j=1,L
        si=xi(j,1)+dxi(j)*(i-1)
330     x(i)=x(i)*si
340     x(i)=x(i)*s
        dx=(x(101)-x(1))/100.
        go to 400
350     xmin=s
        xmax=s
        do 360 i=1,L
        xmin=xmin*xi(i,1)
        m1=k1(i)
360     xmax=xmax*xi(i,m1)
        xmax=alog(xmax)
        xmin=alog(xmin)
        dx=(xmax-xmin)/100.
        do 370 i=1,102
        x(i)=xmin+dx*(i-1)
370     x(i)=exp(x(i))
```

```
400       write( * ,410) m1
410       format(4x,'x(',i3,')')
          write( * ,420) (x(i),i=1,101)
420       format(//5x,9f8.2/(5x,9f8.2))
          do 430 i=1,101
430       nx(i)=0
          x(1)=x(1)-dx/2.
          do 480 i=1,km
          q=s
          do 450 k=1,L
          Ly=la * nyy+bt(1)
          nyy=mod(ly,md)
          rt=float(nyy)/float(md)
          m1=k1(k)
          do 440 j=2,m1
          if(afi(k,j).ge.rt) go to 440
          x1(k)=(xi(k,j)-xi(k,j-1)) * (rt-afi(k,j-1))/
     #    (afi(k,j)-afi(k,j-1))+xi(k,j-1)
          go to 450
440       continue
450       q=q * x1(k)
          do 460 j=1,101
          if(q.gt.x(j).and.q.le.x(j+1)) go to 470
460       continue
          go to 480
470       nx(j)=nx(j)+1
480       continue
          x(1)=x(1)+dx/2.
          s=0.0
          do 500 i=1,101
          p(i)=float(nx(i))/float(km)
          af(i)=1.0-s
500       s=s+p(i)
          write( * ,510) (af(i),i=1,101)
510       format(//4x,3haf=,10f6.4/(7x,10f6.4))
          im=0
          in=0
          do 520 i=1,101
          if(af(i).gt.0.999999) im=im+1
          if(af(i).lt.0.000001) in=in+1
```

```
520     continue
        if(im.le.1.and.in.le.1) go to 590
        dx=100./float(102-im-in)
        write(*,530) im,in,dx
530     format(//2x,3him=,i3,5x,3hin=,i3,5x,3hdx=,f10.4)
        do 540 i=1,101
        xm=float(i-1)+0.000001
        k=ifix(xm/dx)
        xn=xm-dx*k
        j=im+k+1
540     w(i)=(x(j)-x(j-1))*xn/dx+x(j-1)
        do 560 i=1,101
        do 550 j=2,100
        if(x(j).lt.w(i)) go to 550
        go to 560
550     continue
560     r(i)=(af(j)-af(j-1))*(w(i)-x(j-1))/(x(j)-x(j-1))+af(j-1)
        do 570 i=1,102
570     x(i)=w(i)
        do 580 i=1,101
580     af(i)=r(i)
590     do 600 i=1,101
600     f(i)=2*af(i)
        x0=x0-xl-3.
        call map(x,f,101,0)
        y0=y0+3.*f(1)
        return
        end

        subroutine add(x,af,sx,saf,ii,kk,md,la,ny,km)
        dimension x(102),af(101),sx(102),saf(101),ssaf(101)
        common xx(200),xi(10,13),afi(10,12),nx(101),
     #  p(101),w(102),r(101),k1(10),x1(10),dxi(10)
        common/xy/x0,y0,xl,nn(2),bt(2)
        write(3,210) (sx(i),i=1,102)
        write(3,210) (x(j),j=1,102)
        do 100 i=1,102
100     w(i)=x(i)+sx(i)
        do 110 i=1,101
110     nx(i)=0
```

```
      dx=(w(101)-w(1))/100.
      w(1)=w(1)-dx/2.0
      do 180 i=1,km
      do 150 k=1,2
      ly=la*ny+bt(2)
      ny=mod(ly,md)
      f=float(ny)/float(md)
      if(k.eq.1) go to 130
      do 120 j=1,101
      if(saf(j).ge.f) go to 120
      xm=(sx(j)-sx(j-1))*(f-saf(j-1))/(saf(j)-saf(j-1))+sx(j-1)
      go to 150
120   continue
      go to 150
130   do 140 j=1,101
      if(af(j).ge.f) go to 140
      xn=(x(j)-x(j-1))*(f-af(j-1))/(af(j)-af(j-1))+x(j-1)
      go to 150
140   continue
150   continue
      xmn=xm+xn
      do 160 j=1,101
      if(xmn.gt.w(j).and.xmn.le.w(j+1)) go to 170
160   continue
      go to 180
170   nx(j)=nx(j)+1
180   continue
      s=0.0
      do 190 i=1,101
      p(i)=float(nx(i))/float(km)
      r(i)=1.0-s
190   s=s+p(i)
      w(1)=w(1)+dx/2.0
      if(ii.ne.kk) go to 350
      write(*,200)
200   format(//100(1h*))
      write(*,210) (w(i),i=1,101)
210   format(//2x,2hsx,1x,10f12.4/(5x,10f12.4))
      write(*,220) (r(i),i=1,101)
220   format(//2x,3hsaf,10f6.4/(5x,10f6.4))
```

```
        im=0
        in=0
        do 230 i=1,101
        if(r(i).gt.0.999999) im=im+1
        if(r(i).lt.0.000001) in=in+1
230     continue
        if(im.le.1.and.in.le.1) go to 300
        dx=100.0/float(102-im-in)
        write(*,240) im,in,dx
240     format(//2x,3him=,i3,5x,3hin=,i3,5x,3hdx=,f8.4)
        do 250 i=1,101
        xm=float(i-1)+0.000001
        k=ifix(xm/dx)
        xn=xm-dx*k
        j=im+k+1
250     x(i)=(w(j)-w(j-1))*xn/dx+w(j-1)
        do 270 i=1,101
        do 260 j=2,101
        if(w(j).lt.x(i)) go to 260
        go to 270
260     continue
270     af(i)=(r(j)-r(j-1))*(x(i)-w(j-1))/(w(j)-w(j-1))+r(j-1)
        do 280 i=1,102
280     sx(i)=x(i)
        do 290 i=1,101
290     saf(i)=af(i)
        write(3,210) (sx(i),i=1,101)
        write(3,220) (saf(i),i=1,101)
        go to 330
300     do 310 i=1,102
310     sx(i)=w(i)
        do 320 i=1,101
320     saf(i)=r(i)
330     do 340 i=1,101
340     ssaf(i)=saf(i)*2
        call map(sx,ssaf,101,0)
        350 do 360 i=1,102
360     sx(i)=w(i)
        return
        end
```

```
      subroutine map(x,af,n,mdd)
      dimension x(102),af(101)
      common/xy/x0,y0,xl,ny(2),bt(2)
      dx=0. 2
      sa=xl/(x(n)-x(1))
      if(mdd. eq. 1) then
      mp=2
      ag=0.
      xll=7. 5 * dx
      yll=0. 5
      else
      mp=20
      ag=30.
      xll=6. 5 * dx
      yll=-7 * dx * sin(ag)
      end if
      do 100 i=1,n
100   x(i)=x(i) * sa
      sx=x0-x(1)
      call movea(sx+x(1),y0)
      call linea(sx+x(n)+1,y0)
      call linea(sx+x(n)+0. 8,y0-0. 1)
      call movea(sx+x(n)+1,y0)
      call linea(sx+x(n)+0. 8,y0+0. 1)
      call movea(sx+x(1),y0+3. )
      call linea(sx+x(1)-0. 1,y0+2. 8)
      call movea(sx+x(1),y0+3. )
      call linea(sx+x(1)+0. 1,y0+2. 8)
      call movea(sx+x(1).y0+3)
      call linea(sx+x(1),y0)
      do 110 i=1,n
110   call linea(sx+x(i),y0+af(i))
      do 120 i=1,n
      call movea(sx+x(i),y0)
      if(mod(i,mp). eq. 1. and. af(i). gt. 0. 001) then
      yy=x(i)/sa
      call linea(sx+x(i),y0-0. 2)
      call numb1(sx+x(i)-xll,y0-yll,dx,ag,yy)
      end if
```

```
120       continue
          aff=0.4
          call movea(sx+x(1),y0)
          do 130 i=1,6
          fa=1.0-(i-1)*0.2
          call movea(sx+x(1)-0.1,y0+2.-aff*(i-1))
          call linca(sx+x(1),y0+2.-aff*(i-1))
130       call numb1(sx+x(1)-2.4,y0+2.-aff*(i-1),dx,0.,fa)
          do 140 i=1,n
140       x(i)=x(i)/sa
          return
          end
```

§6 应用算例

我国某沉积盆地中的一个地质凹陷有三套生油层系。为估算该凹陷的远景石油资源量，可按氯仿沥青法计算。每套生油层系的石油资源量计算公式如下：

$$Q_j = S \cdot H \cdot D \cdot A \cdot K_1 \cdot K_2 \qquad (10\text{-}33)$$

式中 Q_j——每套生油层系的石油资源量；

S——生油岩分布面积；

H——生油岩厚度；

D——生油岩密度；

A——氯仿沥青含量；

K_1——排烃系数；

K_2——聚集系数。

全凹陷的石油资源量计算公式为

$$Q = \sum_{j=1}^{3} Q_j$$

三套生油层系的地质参数见表 10-4。

表 10-4 三套生油岩层系的地质参数表

地质参数 \ 层系		第一套层系	第二套层系	第三套层系
生油岩分布面积 S/km²		14000	7000	3000
生油岩密度 D/(10⁸t/km³)		23	23	23
排烃系数 K_1		0.44	0.48	0.43
聚集系数 K_2		0.111	0.111	0.111
生油岩厚度	数据个数	140	70	30
	取值范围	0.1~1.0	0.1~1.0	0.1~0.5
氯仿沥青含量 A/%	数据个数	37	37	21
	取值范围	0.03~1.74	0.02~2.08	0.03~1.70

在式(10-33)中，生油岩分布面积 S、生油岩密度 D 为地质常数；排烃系数 K_1、聚集系数

K_2 是由地质类比法确定的经验系数；而生油岩厚度 H、氯仿沥青含量 A 则为有一定取值范围的随机变量。

表 10-5　全凹陷石油资源量汇总表

层位 / 石油资源量/10^8t / 概率/%	第一套生油岩 Q_1	第二套生油岩 Q_2	第三套生油岩 Q_3	全凹陷 Q
100	0.5358	0.3206	0.0988	2.9207
95	2.5089	2.0563	0.2950	11.2233
90	3.6495	3.4874	0.4258	15.9204
85	4.7756	4.8753	0.5462	19.7932
80	5.8729	6.3479	0.6686	23.9546
75	6.8345	7.9966	0.7791	27.7634
70	8.1289	10.1081	0.8814	31.3018
65	9.3880	12.7875	1.0046	35.3488
60	11.2046	15.4970	1.1172	39.7212
55	13.2408	18.1056	1.2553	44.8798
50	15.5740	21.1254	1.4063	50.6544
45	18.4746	24.5984	1.5742	56.9070
40	22.1079	29.0008	1.8469	64.1746
35	25.5952	34.9261	2.1640	70.6543
30	31.0119	43.2521	2.4693	77.7917
25	36.9195	53.4797	2.8498	89.0505
20	46.6373	64.4440	3.3595	103.4409
15	60.5541	76.2474	4.2063	119.8079
10	99.1854	90.8671	5.6153	146.6926
5	148.8125	117.9727	11.5845	180.4146
0	273.6408	178.4241	27.8439	330.2972

三套生油岩层系石油资源量 Q_1、Q_2、Q_3 及全凹陷石油资源总量 Q 在各概率下的数据汇总于表 10-5 中。

从表 10-5 中的数据可以发现，石油资源量 Q_1、Q_2、Q_3 累加后，分布函数的曲线形态有向中间收缩的现象。当概率为 100%时：

$$Q = 2.9270(\times 10^8 t)$$

而

$$Q_1 + Q_2 + Q_3 = 0.5358 + 0.3206 + 0.0988 = 0.9552(\times 10^8 t)$$

可见　$Q > Q_1 + Q_2 + Q_3$

当概率为 0%时：

$$Q = 330.2972(\times 10^8 t)$$

而

$$Q_1 + Q_2 + Q_3 = 273.6408 + 178.4241 + 27.8430 = 479.9079(\times 10^8 t)$$

可见　$Q<Q_1+Q_2+Q_3$

由于 Q_1、Q_2、Q_3 及 Q 的分布函数属于偏态分布(图 10-8 中石油资源量坐标轴的区间是变换后的不等间距区间)，大约在概率 20%处，有

$$Q = 103.4409(\times 10^8 t)$$

$$Q_1 + Q_2 + Q_3 = 46.6376 + 64.4440 + 3.3595$$

$$= 114.4411(\times 10^8 t)$$

所以　$Q \approx Q_1 + Q_2 + Q_3$

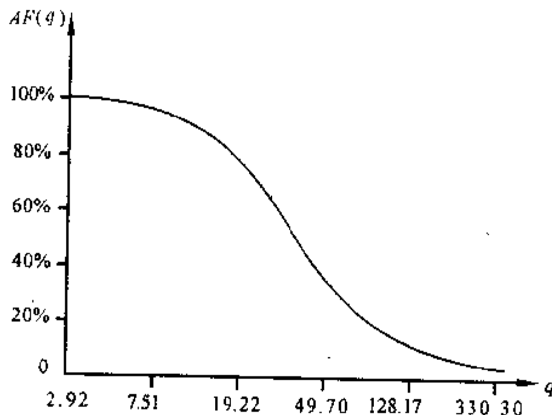


图 10-8　全凹陷的石油资源量分布函数

习　题

1. 蒙特卡罗模拟法的现代含义是什么？
2. 用蒙特卡罗法估算石油资源量时，为什么要产生[0,1]区间上的随机数？
3. 如何构造随机变量的经验分布函数？
4. 在石油资源量模拟计算中如何实现对随机变量的随机抽样？常用抽样方法有几种？
5. 蒙特卡罗法估算石油资源量比传统的估算方法有什么优点？
6. 试述蒙特卡罗法估算局部含油地质单元石油资源量的基本过程。
7. 对求得的石油资源分布曲线 $AF(q)$ 如何解释和应用？

第十一章　盆地模拟简介

§1　盆地模拟的概念

盆地模拟(*Basin Modeling*)是从石油地质的物理化学机理出发，利用地质、地球物理、地球化学、热力学、流体力学、弹性力学等学科的理论知识，将地质人员的概念模型转化为数学模型，最终定量恢复盆地的地质发育史、烃类生成史、运移史和聚集史。盆地模拟是当今石油地质科学领域内为石油勘探服务的的一门新兴课题和技术。从学科角度说，盆地模拟又称“盆地数值模拟”。从软件角度说，盆地模拟又称“盆地模拟系统”，它是盆地数值模拟基础上的整套软件产品。

石油和天然气以流体方式深埋地下，控制它们分布的自然条件十分复杂。究竟是哪些条件控制油气资源的分布？应该怎样去寻找油气藏？这是石油地质学所要回答的主要问题。盆地模拟的主要任务就是实现这个石油地质研究过程的定量化和计算机化，为石油地质研究人员提供一个快速、定量、综合的有效研究手段。盆地模拟是由地史、热史、生烃史、排烃史和运移聚集史等五个模型有机组合的统一体，其内容几乎涉及到石油地质领域的各个分支，这种定量的历史模拟能够直接揭示盆地油气规律的本质，不仅可以从根本上改进与完善石油地质的研究方法，而且能以计算机自动绘图替代繁重的手工绘图。从定量研究的角度，可以说盆地模拟是石油地质领域内的一场革命。

今天，石油地质领域的定量研究及相应的计算机应用仍然是相当薄弱和落后的，而盆地模拟这一新技术正在逐步扭转这一落后局面，使石油地质工作朝着定量化、计算机化和绘图自动化方向发展。多年的实践证明，盆地模拟是实现石油地质定量化的重要途径。盆地模拟不仅仅能对一个完整盆地进行模拟，只要是一个基本独立的油气生聚单元，就可以进行模拟。所以，模拟区域可以是盆地，也可以是坳陷、凹陷或次凹等。

盆地模拟从所研究的空间角度上可分为一维、二维和三维模拟。简单说，一维盆地模拟系统只考虑垂直方向(z 轴)，由平面上各点的模拟结果来综合研究整个模拟地区。用一维盆地模拟可以模拟油气的生成史、初次运移(排烃)史、地层发育史、热史、生油岩成熟史等。但一维盆地模拟的最大缺点是难以模拟油气的二次运移史和油气的聚集史。研究油气的二次运移史的聚集史必须用二维或三维盆地模拟。二维盆地模拟系统考虑的是一个面(z 轴和 x 轴或 r 轴和 y 轴)，它除了能研究一维模拟所能研究的一切问题外，还可以在一个剖面上或平面上研究油气的二次运移史。三维盆地模拟系统考虑的是空间(x、y、z)，所有的数学模型均是三维的。它除了能研究一维和二维模拟所能研究的一切问题外，还可以在空间上研究油气的二次运移史和聚集史。而这在石油勘探中往往是一个很关键的问题。

二维和三维盆地模拟由于受地下情况复杂性的影响，导致难以建立合理描述地下实际情况的数学模型，因此，模拟结果的可信程度必然受到影响。从目前盆地模拟研究的现状看，一维模拟基本上是成熟的、可用的，而多维模拟中的油气运聚模拟基本上是不成熟的、试验性的。多维模拟是否成功的关键在于寻求合理的地质概念模型，并在此基础上建立相应的数学模型。另

外，计算机资源也是限制多维模拟研究的一个因素。因此，无论从实用角度还是从软件角度来说，目前较成熟和较普及的仍然是一维盆地模拟系统。

§2 盆地模拟的发展简史

盆地模拟是当今世界石油地质领域内一个大型综合性新课题。它的发展到目前也只有二十几年的历史。德国、法国、美国、英国、日本、中国是该课题研究水平较高的国家。

早在1969年，当时西德的Tissot B. P等人首先根据化学动力学定律来模拟油源岩中的油气生成量。1978年，西德尤利希核能研究有限公司石油与有机地球化学研究所Yükler M. A和Welte D. H等人建立了世界上第一个一维盆地模拟系统。他们利用求出的古地温史和埋藏史，研究生油层的成熟史，在此基础上进行生烃和排烃的计算。特别值得一提的是，在模拟过程中首次提出了适合求解欠压实地层孔隙流体压力的超压方程。

1984年，法国石油研究院Ungerer. P等人建立了一个较完整的二维盆地模拟系统，除了研究油气的生排历史外，还通过二相运移法研究油气的运移聚集史以及通过地球热力学法求出沿通道运移的含溶解气的油量。1984年美国南卡罗拉那大学地质科学系提出了用镜煤反射率确定古热流的方法，打破了前单纯使用地球热力学法的传统。1988年又提出了用其它几种地化资料确定古热流的方法，使地层古地温史研究的可靠性大为提高。1987年，英国的不列颠石油公司提出了一个关于油气二次运移聚集的二维模型，该模型被认为是目前公开发表的论述油气二次运移与圈闭问题的较好模型之一，研究的基础仍然是达西定律。1981年日本石油勘探有限公司勘探部中山一夫(Nakayama. K)建立了一个简化的二维盆地模拟系统，1988年又建立了一个较完整的二维盆地模拟系统，该系统在烃类生成和运移模型研究上有一定的特色。

我国在1980～1984年，山东胜利油田王捷、韩玉芨等人在和Welter D. H等人合作的基础上，改进了Welte和Yükler的模拟模型，并在山东临邑盆地以及我国东、西部及海上共十几个盆地进行了一维模拟评价，取得了良好效益。同时还对临邑盆地进行了三维模拟。是我国最早开展盆地模拟工作的研究人员。1995年，石油大学查明、张一伟等人运用流体势对油气运聚史进行了二维模拟研究，模拟出了压实流盆地平面流体势历史，通过对东营凹陷的模拟实践，显示模拟结果和油气分布具有良好的关系，为进一步研究油气运聚史开拓了思路。

北京石油勘探开发科学研究院的石广仁、郭秋麟、李惠芬等人，在研究国内外盆地模拟技术的基础上，独立推出了一维盆地模拟系统BAS1，进一步又推出了二维盆地模拟系统BAS2，该系统的特点是采用一维回剥技术(反演法)模拟盆地的沉积史和构造史，用地球热力学和地球化学结合的方法研究古地温史，用热降解法模拟盆地的生烃史，用压实、压差法、渗流力学法计算排烃史。该系统中使用的回剥技术(反演法)别具特色。目前，该系统在石油系统中有一定程度的普及。

上述胜利油田和北京石油勘探开发科学研究院的石广仁的模拟系统，构成了我国盆地模拟研究的主体。另外还有许多单位进行了盆地模拟研究工作，但尚未形成鼎足之势。

§3 盆地模拟的主要模型

盆地模拟由五大模型有机组成：地史、热史、生烃史、排烃史和运移聚集史模型。

一、地史模型

功能：用于描述和重建含油气盆地的沉积史和构造史。

地史模型是盆地模拟技术中的基础模型，其模拟精度直接影响到其他模型模拟的精度。地史模型的主要考虑因素：沉积压实、地层超压、剥蚀、沉积间断、断层等。由于断层因素的复杂多样性，在目前的地史模型中一般只考虑前四种因素。地史模型是盆地模拟的一个基础模型，其可靠程度直接影响到其它模型的运算精度。

方法：分为正演方法和反演方法。

正演方法：从古到今的方法。根据地层的现今厚度及孔隙度等资料，恢复地层的古沉积厚度，后按给定的模型模拟地层从古到今的厚度变化情况。期间可综合考虑地层剥蚀、沉积间断等地质事件，该方法的最大缺点是需要不断调整古地质参数，以使模拟结果与现今的地质资料吻合，如图 11-1。

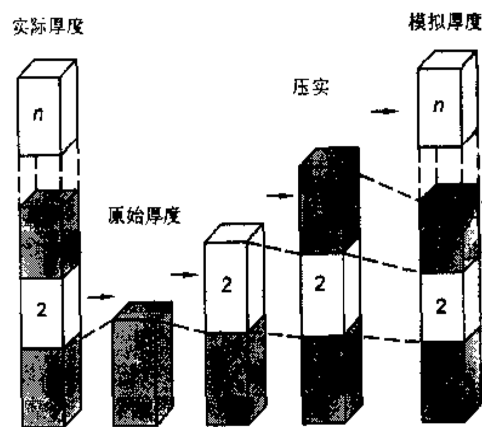


图 11-1　正演法模拟示意图(据查明，1992)

反演方法(回剥技术)：从今到古的方法。从现今的实际地质资料入手，依次恢复地层从今到古的厚度变化情况，相当于把地层从新到老逐层剥去，故称回剥技术。该方法的优点是不需要对模拟结果进行检验，不足之处是不能方便地解决地层超压的问题。

无论是正演还是反演，两种方法都是基于沉积压实原理，即假设随埋藏深度的增加，只有孔隙体积的变小，而地层的骨架厚度不变。

二、热史模型

功能：描述和重建含油气盆地的古热流史和古地温史，进而重建有机质的热成熟史。该模型是盆地模拟的关键模型之一。

热史模型是盆地模拟中的关键模型，它直接影响到生油层生烃史和排烃史模拟的精度，该模型主要考虑因素：热对流、热传导、基底古大地热流值、孔隙度、镜煤反射率及其它地化资料等。

方法：主要有地球热力学方法、地球热力学与地球化学结合的方法。

地球热力学方法：根据盆地现今的大地热流值及其随地质时间的变化，沉积物、孔隙流体和岩石的热导率以及孔隙度随埋深的变化等，来恢复盆地的热流史和温度史。

地球热力学与地球化学结合的方法：基本思路和上述一致，但确定古大地热流值时使用地球热力学与地球化学结合的方法，被认为是更精确的。

单井(一维正演)温度史模拟过程：与地史模拟同步，在纵向上划分节点，用数值方法求解一维温度史模型，得到每个节点上的温度值，并取分布在各地层内节点温度的平均值表示该地层的温度，最终求得各地层温度史(如图 11-2)。

一般，热史模型主要包括的子模型有：温度史模型(地层温度史)、古热流史模型(古大地热流史)、成熟度史模型(生油层 $R_o\%$、产烃率史)。也有的盆地模拟系统将成熟度史模型归纳到生烃史模型中，这些都不是问题的关键。关键在于建立合理实用的模拟模型。

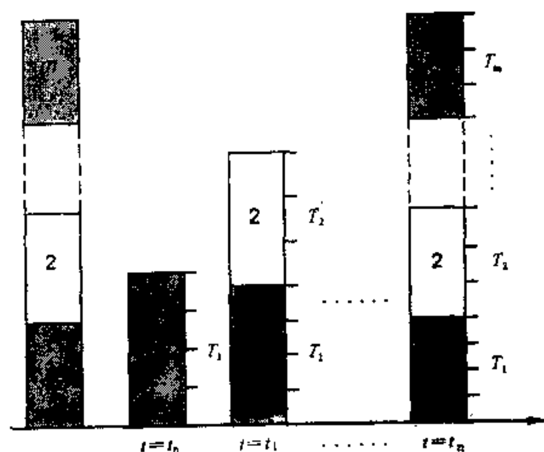


图 11-2 温度史模拟流程示意图

三、生烃史模型

功能：描述和重建含油气盆地的烃类成熟史和生烃量史。

生烃史模型是盆地模拟中的重要模型，其计算结果反映了盆地的生烃能力，并直接影响到排烃史模拟的计算精度。生烃史模型的基础是传统的体积法模型，在此基础上可能演变成计算结果更精确合理的其他形式。

生烃史基本模型：

$$Q=S\cdot H\cdot C_0\cdot \beta\cdot \rho_S$$

式中 Q——生烃量；

S——生油岩面积；

H——生油岩厚度；

C_0——原始有机碳含量；

β——产烃率；

ρ_s——生油岩密度。

若考虑时间 t，则上述模型又可表示为：

$$Q(t)=S\cdot H\cdot C_0\cdot \beta(t)\cdot \rho_S$$

该模型的关键在于产烃率史 $\rho(t)$ 的计算，求解产烃率史常用两种方法：R_o—产烃率关系曲线法（图版法）和化学动力学法。

1. *R_o—产烃率关系曲线法（图版法）*

该方法是在温度史研究的基础上，求出生油岩的 R_o 史，再根据干酪根热模拟实验得出的 R_o—产烃率关系曲线，求出生油层的产烃率史，最终求出生油层的生油量史。

2. *化学动力学法*

Tissot 等人研究认为，干酪根的热降解生油过程遵循化学动力学一级反应定律，干酪根由六类不同键合物质组成，每类键合物质生烃潜力不同，它们按不同的速率降解生油。结合温史研究，利用 Tissot 模型求出各类键合物质各时期的残余量，进而求出降解量，最后求出产烃率。由此计算生油层的生油量史。

综上所述，生烃史模型应该包括以下基本模型：

体积法模型：$Q(t)=S \cdot H \cdot C_0 \cdot \beta(t) \cdot \rho_S$

产烃率史模型：R_o—产烃率图版模型、化学动力学模型

生烃史模型最终的结果是求出各生油层各时期的生油量、生气量。总的说，生烃模型的研究目前是较为成熟的。

四、排烃史模型

功能：描述和重建含油气盆地的排烃历史和排烃方向史（初次运移史）

排烃史模型也是盆地模拟中的重要模型，它反映了生油层的排烃能力。该模型的研究目前还不成熟，主要原因是初次运移的机理复杂多变。现在流行的初次运移的主要机理有：烃类与水呈固有相态运移、水溶液运移、扩散运移、烃溶于气中运移等。一般认为，液态烃的排除主要是基于第一种机理，而气态烃的排除则基于第四种机理。这也是目前盆地模拟系统较多采用的排烃模型设计基础。对于不同的初次运移机理，其排烃模型的设计完全不同。研究排烃史模型的方法从主要有：

压实法（排油）模型～沉积压实排烃

压差法（排油）模型～泥一砂岩之间的压差排烃

渗流力学法（油、气、水）模型～达西定律孔隙流体和超压

五、运移聚集史模型

功能：描述和重建含油气盆地的油气二次运移史和聚集史。包括运移的时间、通道、方向、距离、数量及聚集的位置和数量等。

运移聚集史的研究在盆地模拟中是最困难的，一则地下情况复杂多变，难以准确把握，二则对运聚史的模拟必须要在二维以上的盆地模拟中才能实现，最好是三维模拟，以微型计算机资源是难以承受的。它虽然研究难度大，但却是最重要的部分。目前国内外还没有一个公认的、可靠程度较高的运移聚集史模型。从这点看，运移聚集史模型的研究将是今后盆地模拟技术中的重点研究方向。

目前采用的主要方法有：

二维二相（剖面或平面上油水、气水的二次运移）

三维三相（立体空间中油、气、水的二次运移）

§4 盆地模拟流程及成果输出

一、盆地模拟流程（一维正演）

对一个实际的盆地或地区进行盆地模拟研究时，一般应该经过以下几个步骤：

① 在概念模型的基础上建立合理的数学模型。

② 软件的生成及调试。

③ 模拟网格的划分，形成平面分布的人工井点。

④ 原始资料及模拟参数的输入。

⑤ 对各工人井点模拟运算（从古到今）。

⑥ 模拟结果的检验。

⑦ 模拟结果的输出（数据、图件、磁盘文件）。

⑧ 研究盆地的地史、温度史、生烃史、排烃史及运移史。

上述步骤如图 11-3 所示：

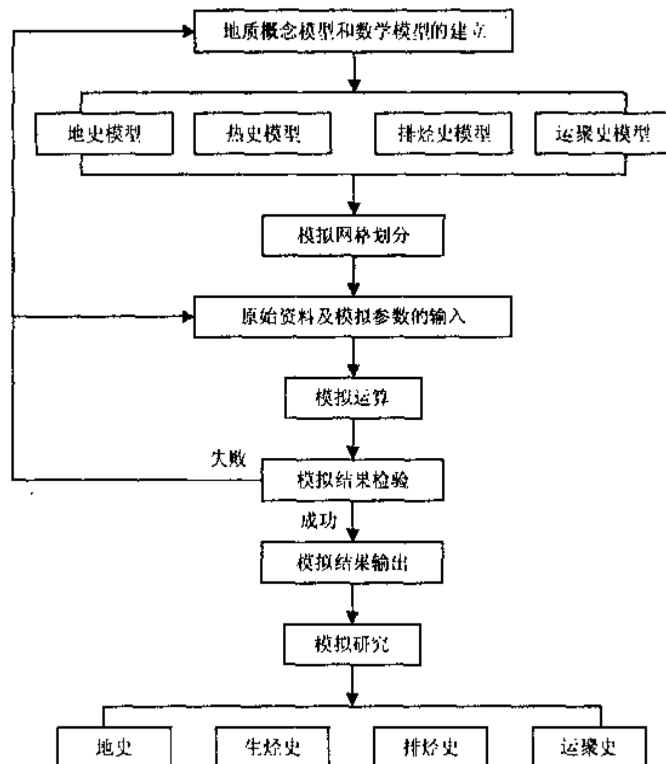


图 11-3 正演法模拟流程图

二、模拟结果输出

1. 输出数据

主要包括系统的地层等厚图数据、孔隙度数据、温度数据、镜煤反射率数据、产烃率数据、生油强度及生油量数据、排烃强度及排烃量数据、排水强度及排水量数据、排液强度及排液量数据、异常压力分布数据等。

2. 输出图件

主要包括模拟平面网格图、各地层各时期的地层等厚图、沉积速率等值图、地层底界埋深等值线图、盆地发育剖面图、单井埋藏史曲线图、古地温等值线图、有机质成熟度图、产烃率分布图、生油强度(单位面积生油量)等值线图、生油剖面图,排烃强度等值线图、异常压力分布图、油气运移方向矢量图、运移速度图、运移流线图、有利地带综合评价图及各种生油层生、排烃量直方图、饼图等统计图形。

§5 盆地模拟发展动向

随着盆地模拟技术的不断发展,预计今后将在以下几个方面取得进展和突破。

一、一维盆地模拟系统的普及

一维盆地模拟系统的使用目前较为成熟,它能进行盆地地史、热史、生烃史、排烃史模拟,但不能进行油气运移聚集史模拟。一维盆地模拟系统以其快速、简捷、实用的特点,尤其适用于勘探程度较低的地区,其使用程度随着微型计算机的普及必将进一步提高。

二、三维系统的出现

研究油气的二次运移和聚集至少要使用二维以上的模拟系统，但油气的运移毕竟是在三维空间进行的，因此，要准确模拟油气的二次运移和聚集，从理论上说以三维模拟系统为最佳。三维模拟系统需要高速大容量的高性能计算机，若三维模拟系统达到了实用水平，能解决解决圈闭的位置及其聚集量等问题，购置高性能的计算机是值得的。

三、地史模型研究

目前在地史模拟中大多仅研究沉积发育史，对于构造发育史的模拟研究不足，今后的盆地模拟系统不仅能处理复杂的断层，而且考虑了盆地的挤压和拉张因素，使构造史的模拟达到实用水平。

四、热史模型研究

将进一步完善地球热力学与地球化学相结合的方法，使热史模拟更加准确。

五、生烃史模型研究

随着地球化学技术的进一步发展，将进一步提高生烃史的计算精度。碳酸岩生烃研究中的问题将得到解决。

六、排烃史模型研究

随排烃机理研究进一步深入以及三维模拟系统的使用，排烃史模型将更加完善。

七、运移聚集史模型

随着构造史模拟的完善以及三维混相（甚至多组分）渗流力学方法的引入，将能够较正确地指出圈闭位置及其聚集量。

八、盆地模拟数据库

包括模拟参数数据库和模拟结果数据库的建立，将更有利于资料的查询、统计、输出等数据管理和资源共享。

九、模拟系统的集成化

盆地模拟系统将与遥感信息处理系统、地震资料处理系统、测井资料解释系统等建立良好的通讯接口，实现快速、实用的一体化盆地模拟系统。

十、绘图技术

将进一步完善盆地模拟系统的图形输入输出系统，传统的手工绘图将被计算机自动绘图所取代。

习　题

1. 什么是盆地数值模拟？盆地数值模拟的主要任务是什么？
2. 盆地数值模拟的主要模型包括哪些？它们的作用是什么？
3. 叙述盆地数值模拟的工作流程。

第十二章　模拟模型

§1　地史模型

地史模型的功能是描述和重建含油气盆地的沉积发育史和构造史。其作用在于为热史、生烃史、排烃史、运聚史提供时空模拟范围。油气藏的形成是发生在地史过程中的事件，在沉积物的埋藏、成岩过程中，其密度、孔隙度、厚度、结构均发生变化，这种变化对油气藏的形成是至关重要的。

地史模型是盆地模拟的基础模型。其精度直接影响到其余的四个模型：热史、生烃史、排烃史、运聚史模型，并为它们提供有关参数。地史模拟过程中，应考虑尽可能多的地质事件，如沉积压实、超压、剥蚀、沉积间断、断层等。地史模型目前的主要应用是对盆地沉积发育史的模拟，对包含断层发育史在内的构造发育史的模拟难度较大，目前尚不成熟。本章主要讨论盆地沉积发育史(埋藏史)的模拟。

一、概念模型

盆地的沉积埋藏史模拟主要是基于沉积地层的压实原理实现的。根据沉积压实原理，假设地层随着埋藏深度的增加，只有孔隙体积变小，而的"骨架"厚度不变，符合这一原理的主要是砂、泥(页)岩类，而碳酸盐岩、塑性流动的膏盐层、火山岩等因成岩作用机理不同，在模拟时要特别对待。因此，相应数学模型主要是针对碎屑岩类发育的盆地而建立的。其概念模型的建立主要从以下几个方面考虑：

① 沉积地层厚度及其变化，既反映了上覆沉积对下伏地层的压力效应，又反映了不同岩石因受压实程度不同所引起的孔隙度非均匀变化，因此根据压实原理，用现今地层厚度和孔隙度可以恢复地层的原始厚度。

② 地层被抬升、剥蚀是盆地发展过程中重要事件，抬升时间和剥蚀量则是恢复盆地发展演化史的两个重要参数。用适当的方法确定这两个地质参数，并将其与原始地层厚度一起考虑进行地史模拟，可以恢复盆地的沉积埋藏史。

③ 多种原因形成的地层欠压实作用(超压带的存在)是较为普遍的地质现象，此时因孔隙度的变化不再遵循 Athy 定律，恢复的地层厚度与真正的原始厚度有差异，概念模型必须考虑这一因素。

另外，构造变形和断裂作用、次生孔隙等因素，都为沉积埋藏恢复增加了许多难度，目前的模拟方法仍难以处理。

基于沉积压实原理上的地史模型分为正演法模型和反演法模型。以下分别进行介绍。

二、正演法模型

建立地史模型的目的是模拟研究盆地的沉积发育史和构造史，由于构造史模拟的不成熟性，一般只模拟沉积发育史，换句话说，就是模拟研究盆地或模拟地区内所沉积的各套地层的厚度变化历史。要达到这个目的，我们首先要解决以下三个问题：

① 求解各地质时期地层孔隙流体压力，即地层压力史。

② 求解各地质时期地层孔隙度，即地层孔隙度史。

③ 求解各地质时期地层厚度，即地层厚度史。

基于上述三个问题，地史模型一般由以下三个子模型构成。

① 压力史模型。

② 孔隙度史模型。

③ 地层厚度恢复模型。

1. 压力史模型

地层在沉积过程中，由最初的沉积厚度演变到今天的沉积厚度，其中经过了一个压实、即孔隙度减小的过程。我们认为，孔隙度的减小，主要依赖于地层上覆沉积荷重和孔隙流体压力的变化。当知道地层所处的深度和上覆岩石的平均密度后，上覆沉积荷重不难由下列公式求出：

$$S = \rho_{bw} \cdot g \cdot Z \tag{12-1}$$

式中 S——上覆沉积荷重；

ρ_{bw}——上覆沉积物平均密度；

g——重力加速度；

Z——埋深。

上覆沉积荷重可用上式求解。另外一个关键问题是如何求取地层的孔隙流体压力。在静水压力条件下的正常压实地层，孔隙流体压力 P_n 可由下列公式求得(据真炳钦次，1968)：

$$P_n = \rho_w \cdot g \cdot Z \tag{12-2}$$

式中 P_n—— 孔隙流体压力(地静压力)；

ρ_w——孔隙流体密度；

g——重力加速度；

Z——埋深。

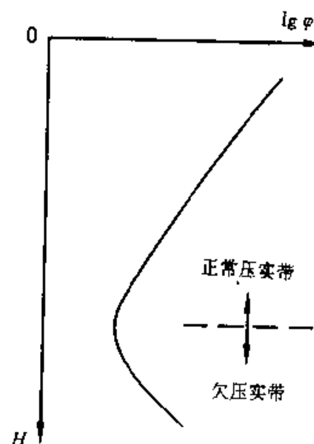


图 12-1 有限压实的孔隙度-深度曲线

在盆地模拟过程中，如果从下到上的所有地层中均未出现欠压实现象，即均为静水压力条件下的正常压实地层，我们就可以用式(12-2)求解任何深度的孔隙流体压力，压力史模型的问题就基本解决了。但在实际的沉积地层中，往往在下部地层中出现欠压实现象，即孔隙度不但不随深度的增大而减小，反而有增大的趋势。如图 12-1 所示。在这种情况下，对孔隙流体压力的求解，Magara. K (真炳钦次)曾介绍过一种方法。

(1) Magara 欠压实带孔隙流体压力方程

Magara. K 给出了一个计算异常压力的公式：

$$P = \rho_w \cdot g \cdot Z_e + \rho_{bw} \cdot g \cdot (Z - Z_e) \tag{12-3}$$

或 P(磅/英寸^2) $= \gamma_w \cdot Z_e + \gamma_{bw} \cdot (Z$(英尺)$- Z_e$(英尺))

式中 P——深度 Z 处的孔隙流体压力(含超压)；

Z_e——正常压实趋势上某一较浅深度，该处的孔隙度等于深 Z 处岩石的孔隙度；

ρ_w——孔隙流体平均密度；

ρ_{bw}——沉积物平均密度；

g——重力加速度；

γ_w——孔隙流体平均密度，磅/英寸^2·英尺；

γ_{bw}——沉积物平均密度，磅/英寸^2·英尺。

Z_e 和 Z 的关系如图 12-2 所示。

下面我们来推导真柄的公式。

根据特察模型知：

$$S=\sigma+P$$

式中 S——上覆沉积荷重；

σ——作用在岩石上的有效应力（颗粒间的支撑力）；

P——孔隙流体压力。

在深 Z_e处：

$$S_e=\rho_{bw}\cdot g\cdot Z_e$$

$$P_e=\rho_w\cdot g\cdot Z_e$$

$$\sigma_e=S_e-P_e=\rho_{bw}\cdot g\cdot Z_e-\rho_w\cdot g\cdot Z_e$$

在深 Z 处：

$$S=\rho_{bw}\cdot g\cdot Z$$

$$\sigma=S-P=\rho_{bw}\cdot g\cdot Z-P$$

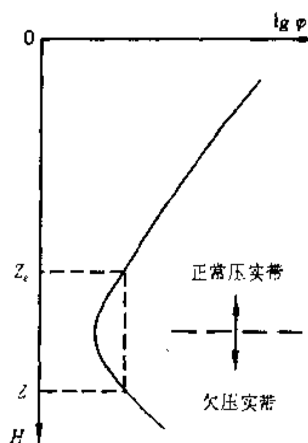


图 12-2 欠压实孔隙度-深度曲线上的 Z 和 Z_e

由于深 Z_e 处和深 Z 处的孔隙度相等，我们可以认为这时作用在这两个深度岩石颗粒间的支撑力也是相同的。

即有：

$$\sigma_e=\sigma$$

$$\rho_{bw}\cdot g\cdot Z_e-\rho_w\cdot g\cdot Z_e=\rho_{bw}\cdot g\cdot Z-P$$

即 $$P=\rho_w\cdot g\cdot Z_e-\rho_{bw}\cdot g\cdot Z_e+\rho_{bw}\cdot g\cdot Z=\rho_w\cdot g\cdot Z_e+\rho_{bw}\cdot g\cdot(Z-Z_e)$$

即 $$P=\rho_w\cdot g\cdot Z_e+\rho_{bw}\cdot g\cdot(Z-Z_e)$$

正是式(12-3)。

用式(12-3)可求出在任一地质时间和深度的孔隙流体压力，但该公式实际应用的效果并不理想，甚至往往出现很大的误差，这主要是因为异常流体压力的起因并不象上述方程所反映的那样，仅仅是由于上覆沉积物的不断加厚所至所致。另外诸如岩石渗透率、水热增压等重要因素均未考虑在内。这一点我们可以从特察模型中可以看出。由特察模型：$S=\sigma+P$，岩石在从 Z_e沉积到 Z 的过程中，岩石骨架的有效应力 σ 不变，孔隙流体压力的变化完全是由于上覆沉积荷重 S 的增加，这样考虑问题显然不够全面。

由于真柄公式忽略了造成异常压力的其他主要因素，故所计算出的流体压力值往往小于实测值。一个解决的办法是在其公式后加一修正因子 α，即：

$$P=\rho_w\cdot g\cdot Z_e+\alpha\cdot\rho_{bw}\cdot g\cdot(Z-Z_e) \qquad (12\text{-}4)$$

或 $$P=P_n-\alpha\cdot P_a$$

其中 α 的值可根据现今的实测资料与原公式的对比修正计算得到。虽如此，上述公式的计算结果仍不十分理想，因此，我们有必要寻求求解欠压实带孔隙流体压力的更好方法。

(2) 欠压实带孔隙流体压力超压方程及其推导

胜利油田王捷、韩玉茇等人在 Welte 模型的基础上，曾推导了一个求解欠压实带孔隙流体

压力的三维偏微分方程，即超压方程。

$$\nabla(\frac{\rho_w K}{\mu}\nabla P_a) = [(1-\varphi)\alpha + \varphi\beta]\rho_w \frac{\partial P_a}{\partial t} - \alpha(1-\varphi)\rho_w \frac{\partial S}{\partial t} + [(1-\varphi)\alpha + \varphi\beta]\rho_w \frac{\partial P_n}{\partial t} \quad (12\text{-}5)$$

式中 ∇——算子，$\nabla = \frac{\partial}{\partial x}\vec{i} + \frac{\partial}{\partial y}\vec{j} + \frac{\partial}{\partial z}\vec{k}$；

ρ_w——孔隙流体密度；

μ——流体粘度；

K——岩石渗透率；

P_a——孔隙流体超压；

φ——岩石孔隙度；

α——岩石压缩系数；

β——流体压缩系数；

S——上覆沉积荷重；

P_n——上覆静水柱压力；

t——时间。

下面来推导超压方程。我们知道，地下的岩石是由岩石骨架和孔隙流体组成，据特察(Terzaghi. K)1968年的实验，对任意一块岩石，作用在它上面的有效应力 σ 等于作用其上的上覆荷重 S 与孔隙流体压力 P 之差，即：

$$\sigma = S - P \quad (12\text{-}6)$$

另外，若以 V 表示岩块体积，V_s 表示岩石骨架体积，φ 表示岩石孔隙度，则有：

$$V_s = (1-\varphi)\cdot V \quad (12\text{-}7)$$

由于我们假设岩石骨架是不可压缩的，因此含有孔隙的沉积物的压缩完全体现在因沉积物中流体排出而引起的孔隙空间的变小，其骨架体积 V_s 不随时间 t 而变化。即有：

$$\frac{\partial V_s}{\partial t} = 0$$

将式(12-7)代入上式得：

$$\frac{\partial V_s}{\partial t} = -V\frac{\partial \varphi}{\partial t} + (1-\varphi)\frac{\partial V}{\partial t} = 0$$

即

$$\frac{\partial \varphi}{\partial t} = \frac{(1-\varphi)}{V}\cdot\frac{\partial V}{\partial t} \quad (12\text{-}8)$$

由虎克定律知：岩块的相对形变与作用其上的有效应力 σ 的变化成正比，即：

$$\frac{\Delta V}{V} = -\alpha\cdot\Delta\sigma$$

式中 α——岩石压缩系数(指岩石压实过程中体积的变化率，反映了岩石的可压缩程度)。

将上式两边同除 Δt 得：

$$\frac{1}{V}\cdot\frac{\Delta V}{\Delta t} = -\alpha\frac{\Delta\sigma}{\Delta t}$$

用对时间的偏导数可将上式表示为：

$$\frac{1}{V}\cdot\frac{\partial V}{\partial t} = -\alpha\frac{\partial\sigma}{\partial t}$$

结合式(12-8)有：

$$\frac{\partial\varphi}{\partial t} = -\alpha(1-\varphi)\frac{\partial\sigma}{\partial t} \tag{12-9}$$

将式(12-6)代入式(12-9)得：

$$\frac{\partial\varphi}{\partial t} = -\alpha(1-\varphi)\left(\frac{\partial S}{\partial t} - \frac{\partial P}{\partial t}\right) \tag{12-10}$$

式(12-10)即为计算孔隙度 φ 随 S 和 P 的变化而变化的基本公式。

我们再假设：岩块内的流体是可以压缩的。对可压缩的流体，虎克定律描述了流体体积的相对变化与孔隙流体压力的变化成正比，即：

$$\frac{\mathrm{d}V_w}{V_w} = -\beta \cdot \mathrm{d}P \tag{12-11}$$

式中　V_w——岩石中流体体积；

β——流体压缩系数；

P——孔隙流体压力。

假定当孔隙流体压力从 P_0 变化到 P 时，相应的流体体积由 V_{w_0} 变化到 V_w，据此对上式两边取定积分：

$$\int_{V_{w0}}^{V_w} \frac{\mathrm{d}V_w}{V_w} = -\int_{P_0}^{P} \beta \mathrm{d}P$$

由此可知

$$\ln V_w - \ln V_{w0} = \beta(P_0 - P)$$

即

$$\ln\left(\frac{V_w}{V_{w0}}\right) = \beta(P_0 - P)$$

$$\frac{V_w}{V_{w0}} = \exp[\beta(P_0 - P)]$$

即有

$$V_w = V_{w0} \cdot \exp[\beta(P_0 - P)] \tag{12-12}$$

设 M 是体积为 V_w 时的流体质量，ρ_w 是体积为 V_w 时的流体密度，则有：

$$\rho_w = \frac{M}{V_w} = \frac{M}{V_{w0}}\exp[-\beta(P_0 - P)]$$

$$= \rho_{w0} \cdot \exp[-\beta(P_0 - P)] \tag{12-13}$$

式中　ρ_{w0}——体积为 V_{w0} 时的流体密度。

将上式两边对时间求偏导可得：

$$\frac{\partial\rho_w}{\partial t} = \rho_{w0} \cdot \exp[-\beta(P - P_0)]\frac{\partial P}{\partial t} \cdot \beta$$

整理可得：

$$\frac{\partial\rho_w}{\partial t} = \rho_w \cdot \beta\frac{\partial P}{\partial t} \tag{12-14}$$

该式表明流体密度的变化率与孔隙流体压力的变化率成正比。

另外，描述质量守恒的连续性方程为：

$$-\nabla(\rho_w\vec{V}) = \frac{\partial(\rho_w\varphi)}{\partial t} \tag{12-15}$$

式中　$\vec{V}$——速度向量；

∇——算子，$\nabla = \frac{\partial}{\partial x}\vec{i} + \frac{\partial}{\partial y}\vec{j} + \frac{\partial}{\partial z}\vec{k}$。

由式(12-10)和式(12-14)知：

$$\frac{\partial\varphi}{\partial t}=-\alpha(1-\varphi)\cdot\frac{\partial\sigma}{\partial t}$$

$$\frac{\partial\rho_w}{\partial t}=\rho_w\cdot\beta\frac{\partial P}{\partial t}$$

代人式(12-15)右端得：

$$\frac{\partial(\rho_w\varphi)}{\partial t}=\varphi\rho_w\beta\frac{\partial P}{\partial t}-\rho_w\alpha(1-\varphi)\frac{\partial\sigma}{\partial t} \tag{12-16}$$

另外，重力作用下渗流的达西定律可表示为：

$$\vec{V}=-\frac{K}{\mu}\nabla(P-\rho_w\cdot g\cdot Z) \tag{12-17}$$

式中 K——渗透率；

μ——粘度；

g——重力加速度；

Z——距水平面的深度。

将式(12-17)代人式(12-15)并结合式(12-16)得：

$$\nabla\left[\frac{\rho_w K}{\mu}\nabla(P-\rho_w\cdot g\cdot Z)\right]=\varphi\rho_w\beta\frac{\partial P}{\partial t}-\rho_w\alpha(1-\varphi)\frac{\partial\sigma}{\partial t} \tag{12-18}$$

设 P 为孔隙流体压力，P_n为静水柱压力，P_a为孔隙流体超压，由于 $P_n=\rho_w\cdot g\cdot Z$，而 $P=P_n+P_a$，故有：

$$\begin{aligned}P_a&=P-P_n\\&=P-\rho_w\cdot g\cdot Z\end{aligned} \tag{12-19}$$

将式(12-19)及式(12-6)代人式(12-18)并整理得：

$$\nabla\left(\frac{\rho_w K}{\mu}\nabla P_a\right)=[(1-\varphi)\alpha+\varphi\beta]\rho_w\frac{\partial P_a}{\partial t}-\alpha(1-\varphi)\rho_w\frac{\partial S}{\partial t}+[(1-\varphi)\alpha+\varphi\beta]\rho_w\frac{\partial P_n}{\partial t}$$

超压方程推导完毕。在地史模拟的过程中，正是以上述方程来求解沉积地层内的孔隙流体压力的变化情况。超压方程的提出，为地史模型的建立提供了基础。

综上所述，一个完整的压力史模型为：

$$\begin{cases}P=P_r+P_a\\ \text{超压力方程求 } P_a\end{cases}$$

2. 孔隙度史模型

在推导超压方程的过程中，我们推导出了式(12-10)：

$$\frac{\partial\varphi}{\partial t}=-\alpha(1-\varphi)\left(\frac{\partial S}{\partial t}-\frac{\partial P}{\partial t}\right)$$

上式即是地史模型中使用的孔隙度史模型。从该式中可以看出，地层孔隙度随时间的变化，依赖于上覆沉积荷重 S 随时间的变化以及孔隙流体压力随时间的变化。

若设地层中垂向某一点从时间 t_i 演化到 t_{i+1}，孔隙度从 $\varphi(t_i)$变化到 $\varphi(t_{i+1})$，上覆沉积荷重 S 从 $S(t_i)$变化到$S(t_{i+1})$，P 从 $P(t_i)$变化到 $P(t_{i+1})$，如果时间间隔足够小，则根据上述孔隙度史模型，近似有：

$$\varphi(t_i)-\varphi(t_{i+1})=-\alpha[1-\varphi(t_i)][(S(t_i)-S(t_{i+1})-P(t_i)+P(t_{i+1})]$$

即 $$\varphi(t_{i+1}) = \varphi(t_i) + \alpha[1 - \varphi(t_i)][(S(t_i) - S(t_{i+1}) - P(t_i) + P(t_{i+1})] \quad (12\text{-}20)$$

在实际的地史模拟过程中，就是根据这个公式依次求出地层孔隙度的变化历史。若设地层经历的地质时刻依次为：t_0、t_1、t_2、…t_n，求解孔隙度史的过程如下。

当 $t=t_0$ 时：

① 求地层的孔隙流体压力 $P(t_0)$。

② 求地层的上覆沉积荷重 $S(t_0)$。

③ 确定地层的原始孔隙度 $\varphi(t_0)$。

当 $t=t_1$ 时：

① 由压力史模型，求出地层的孔隙流体压力 $P(t_1)$。

② 求地层的上覆沉积荷重 $S(t_1)$。

③ 由式(12-20)求地层的孔隙度 $\varphi(t_1)$。

当 $t=t_2$ 时：

① 由压力史模型，求出地层的孔隙流体压力 $P(t_2)$。

② 求地层的上覆沉积荷重 $S(t_2)$。

③ 由式(12-20)求地层的孔隙度 $\varphi(t_2)$。

⋮ ⋮ ⋮

依次类推。

当 $t=t_n$时：

① 由压力史模型，求出地层的孔隙流体压力 $P(t_n)$。

② 求地层的上覆沉积荷重 $S(t_n)$。

③ 由式(12-20)求地层的孔隙度 $\varphi(t_n)$。

最终即可求得地层在时刻 t_0、t_1、t_2、…、t_n时的孔隙度 $\varphi(t_0)$、$\varphi(t_1)$、$\varphi(t_2)$、…，$\varphi(t_n)$，即求出了孔隙度的变化历史。

3. *地层厚度恢复模型*

基本模型：
$$H_0=\frac{1-\varphi_p}{1-\varphi_0}H_p \quad (12\text{-}21)$$

式中 H_0——地层的原始厚度；

H_p——地层的现今厚度(已知)；

φ_0——地层的原始孔隙度(已知)；

φ_p——地层的现今孔隙度(已知)。

若根据孔隙度史模型求出了地层在 t_i时刻的孔隙度 $\varphi(t_i)$，$i=1,2,\cdots,n$，则该时刻的地层厚度 $H(t_i)$可由下式求出：

$$H(t_i) = \frac{1-\varphi_0}{1-\varphi(t_i)}H_0 \qquad (i = 1,2,\cdots,n) \quad (12\text{-}22)$$

据此可求出地层在时刻 t_0、t_1、t_2、…、t_n 时的厚度 $H(t_0)$、$H(t_1)$、$H(t_2)$、…、$H(t_n)$，即求出了厚度的变化历史。

地层的厚度史求出之后，就达到了地史模拟的第一个目的。当我们对在一个沉积盆地或模拟地区内均匀分布的许多人工井点进行这样的模拟后，我们就不难从得到的地层的厚度变化历史去研究整个盆地或模拟地区的沉积发育史。

根据地层的厚度历史，我们可以绘制盆地演化宝塔图(图 12-3)。也可以绘制某剖面的地

层发育剖面图(图 12-4)。

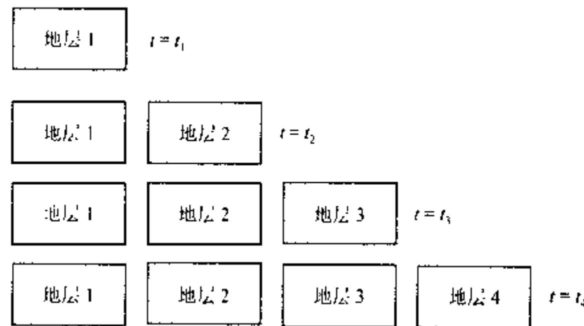


图 12-3 盆地演化宝塔示意图

另外还可以绘制单井埋藏史曲线图等。

4. 模拟时间、空间步长划分及超压史模拟过程(一维)

在一维模拟情况下,我们需要确定时间和空间(垂向)模拟计算步长。步长的确定视盆地的具体情况、模拟的精度要求和计算机资源的不同而定,时间和空间步长取得越小,计算精度越高,所用机时越长。一般,常选取的时间步长有 10、20、50 万年,垂向上的空间步长一般取 10~50m。在选定时间步长内的地层沉积厚度应大于垂向空间步长。

在选定时间步长 Δt 和空间步长 Δz,地层孔隙流体超压史模拟过程大致如下。

① 取第一个时间步长 Δt ,该时间阶段内沉积有一定厚度的地层,根据所确定的垂向空间步长 Δz ,将沉积的地层在垂向上从下到上划分为若干小段(假设 5 小段),这样在垂向上共形成了若干节点(6 个节点),求解每个节点上的超压 P_a^i,$i=1,2,\cdots,6$。若节点 i 处于正常压实带,取 $P_a^i=0$。

② 取第二个时间步长 Δt ,该时间阶段内又沉积有一定厚度的地层,根据空间步长 Δz ,将新沉积的地层在垂向上从下到上划分为若干小段(假设 4 小段),这样在垂向上增加了若干节点(4 个节点),求解已有的每个节点上的超压 P_a^i,$i=1,2,\cdots,10$。若节点 i 处于正常压实带,取 $P_a^i=0$。

③ 取第三个时间步长 Δt ,该时间阶段内又沉积有一定厚度的地层,根据空间步长 Δz ,将新沉积的地层在垂向上从下到上划分为若干小段(假设 3 小段),这样在垂向上又增加了若干节点(3 个节点),求解所有每个节点上的超压 P_a^i,$i=1,2,\cdots,13$。若节点 i 处于正常压实带,取 $P_a^i=0$。

⋮ ⋮ ⋮ ⋮ ⋮ ⋮

依次类推。

④ 取最后一个时间步长 Δt ,该时间阶段内又沉积有一定厚度的地层,根据空间步长 Δz ,将新沉积的地层在垂向上从下到上划分为若干小段(假设 5 小段),这样在垂向上又增加了若干节点(5 个节点),求解所有每个节点上的超压 P_a^i,$i=1,2,\cdots,n$。若节点 i 处于正常压实带,取 $P_a^i=0$。其中 n 为所有节点的总数。

超压史求解完毕。

在模拟过程中,正常压实带和欠压实带的分界线是这样确定的,在模拟之前先确定盆地出

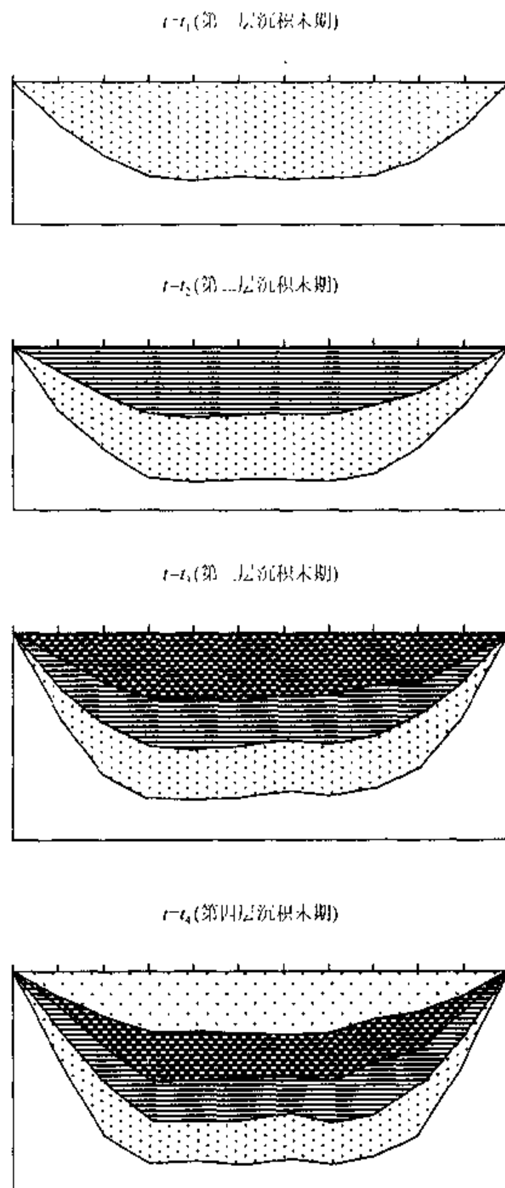


图 12-4 沉积发育剖面示意图

现欠压实的层位，由此可以确定平面上每个模拟点欠压实出现的深度，当节点处于该深度之下时，认为该节点处于欠压实带，否则认为该节点处于正常压实带。

5. 超压方程的适用范围

我们在前面介绍过，建立超压方程的目的是为了求取地层的孔隙流体超压，因此，它只能使用于地层超压存在的区域。对盆地中的所有沉积物而言，在垂向大致上可以划分为三个区域。如图 12-5 所示，最上面是水域，中间是正常压实区域，该区域沉积岩内的孔隙流体压力为静水柱压力，孔隙流体超压 $P_a=0$。最下部为欠压实区域，处在该区域的地层无论其岩性如何，

一般我们统一认为有超压存在，随着沉积物不断加厚，这个层的上边界相对向上移动，超压方程仅仅适用于这个区域。

6. 超压方程的数值解法

求解超压方程一般使用有限差分等数值解法，由于二维和三维模拟在解方程时方法繁琐，而且就一般的微机而言，进行多维模拟的能力也不够。因此，本节中我们仅就一维超压方程介绍其有限差分数值解法。

一维超压方程可表示为：

$$\frac{\partial}{\partial z}\left(\frac{\rho_w K}{\mu}\cdot\frac{\partial P_a}{\partial z}\right)=[(1-\varphi)\alpha+\varphi\beta]\rho_w\frac{\partial P_a}{\partial t}-\alpha(1-\varphi)\rho_w\frac{\partial S}{\partial t}+[(1-\varphi)\alpha+\varphi\beta]\rho_w\frac{\partial P_n}{\partial t} \tag{12-23}$$

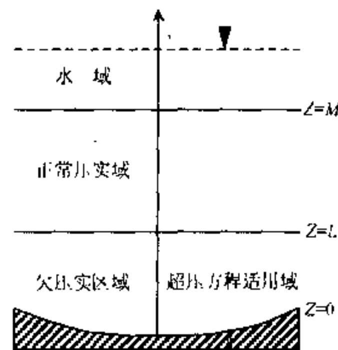


图 12-5 超压方程适用范围

超压方程的边界条件可处理为：考虑盆地的基底为致密岩石组成的不渗透地层，故下边界条件$\frac{\partial P_a}{\partial z}=0$ $(Z=0)$，而上边界为正常压实和欠压实的分界线，故上边界条件 $P_a=0$ $(Z=L)$。

因此，一维超压方程的定解问题归纳为：

$$\begin{cases}\frac{\partial}{\partial z}\left(\frac{\rho_w K}{\mu}\frac{\partial P_a}{\partial z}\right)=[(1-\varphi)\alpha+\varphi\beta]\rho_w\frac{\partial P_a}{\partial t}-\alpha(1-\varphi)\rho_w\frac{\partial S}{\partial t}+[(1-\varphi)\alpha+\varphi\beta]\rho_w\frac{\partial P_n}{\partial t}\\ \frac{\partial P_a}{\partial t}=0 \qquad (Z=0)\\ P_a=0 \qquad (Z=L)\end{cases}$$

按一维模拟所确定的时间、空间步长及垂向空间网格划分方法，一维超压方程的差分格式如下：

$$\Delta Z_k\Delta P_{a2}-E^iP_{a2}^i=F^i-E^iP_{a1}^i \tag{12-24}$$

式中 P_{a1}——前阶段的超压；

P_{a2}——本阶段的超压；

i——空间网格节点的位置；

$\Delta Z_k\Delta P_{a2}$——综合项，$\Delta Z_k\Delta P_{a2}=Z_k^{i+\frac{1}{2}}(P_{a2}^{i+1}-P_{a2}^i)+Z_k^{i-\frac{1}{2}}(P_{a2}^{i-1}-P_{a2}^i)$。

故原差分格式可改写为：

$$Z_k^{i+\frac{1}{2}}P_{a2}^{i+1}-(Z_k^{i+\frac{1}{2}}+Z_k^{i-\frac{1}{2}}+E^i)P_{a2}^i+Z_k^{i-\frac{1}{2}}P_{a2}^{i-1}=F^i-E^iP_{a1}^i \tag{12-25}$$

式中

$$Z_k^{i+\frac{1}{2}}=\frac{4\left(\frac{K\rho_w}{\mu}\right)^{i+1}\left(\frac{K\rho_w}{\mu}\right)^{i}\Delta t_m}{\left[\left(\frac{K\rho_w}{\mu}\right)^{i}+\left(\frac{K\rho_w}{\mu}\right)^{i+1}\right](\Delta Z^{i+1}+\Delta Z^i)}$$

$$Z_k^{i-\frac{1}{2}}=\frac{4\left(\frac{K\rho_w}{\mu}\right)^{i}\left(\frac{K\rho_w}{\mu}\right)^{i-1}\Delta t_m}{\left[\left(\frac{K\rho_w}{\mu}\right)^{i}+\left(\frac{K\rho_w}{\mu}\right)^{i-1}\right](\Delta Z^{i}+\Delta Z^{i-1})}$$

$$E^i=[(1-\varphi)\alpha+\varphi\beta]^i\rho_w^i\Delta Z^i$$

$$F^i=\{[(1-\varphi)\alpha+\varphi\beta][(\rho_w gZ)_2-(\rho_w gZ)_1]-(1-\varphi)\alpha(S_2-S_1)\}^i\rho_w^i\Delta Z^i$$

式中 Δt_m——时间步长，Ma；

ΔZ^i——空间步长，cm；

$(\rho_w gZ)_1$——前阶段静水柱压力；

$(\rho_w gZ)_2$——本阶段静水柱压力；

S_1——前阶段上覆荷重；

S_2——本阶段上覆荷重。

令：　$a_i=Z_k^{i+\frac{1}{2}}, b_i=-(Z_k^{i+\frac{1}{2}}+Z_k^{i-\frac{1}{2}}+E^i), c_i=Z_k^{i-\frac{1}{2}}, f_i=F^i-E^iP_{a1}^i, x_i=P_{a2}^i$

则差分方程(12-25)成为：

$$a_i x_{i+1}+b_i x_i+c_i x_{i-1}=f_i \qquad (i=1,2,\cdots,n) \tag{12-26}$$

该方程是关于 $x\ (P_{a2})$ 的三对角线形方程组，通常采用追赶法求解。

将追赶法的递推关系表示如下：

$$A_i=\frac{f_i-c_iA_{i-1}}{b_i+c_iB_{i-1}} \tag{12-27}$$

$$B_i=\frac{-a_i}{b_i+C_iB_{i-1}} \tag{12-28}$$

$$x_i=A_i+B_ix_{i+1} \tag{12-29}$$

下边界条件$\frac{\partial P_a}{\partial z}=0$处理为 $c_2=Z_k^{2-\frac{1}{2}}=0$，由式(12-27)和式(12-28)得：$A_2=\frac{f_2}{b_2}, B_2=-\frac{a_2}{b_2}$，从而可依次求出：$A_i, B_i(i=2,\cdots,n)$。对上边界 $P_{a2}=0$，若指定上边界之下的第一个内点的流体超压为 P_{a2}^n，则可以认为：$P_{a2}^{n+1}=0$。

因此，按式(12-29)有：$P_{a2}^n=A_n+B_n\cdot P_{a2}^{n+1}=A_n$，即 $P_{a2}^n=A_n$，A_n 前已求得，再按递推公式(12-29)可求出 $P_{a2}^{n-1}, P_{a2}^{n-2}, \cdots, P_{a2}^2$。超压方程求解完毕。

三、地层回剥技术简介

主要思路：各地层在保持其骨架厚度不变的条件下，从今天盆地的分层现状出发，按地质年代从新到老(由今至古)逐层剥去，直至全部剥完。

回剥技术适用于正常压实带，即超压为零的情况，它所用的关键参数是孔隙度—深度曲线。而孔隙度—深度曲线应满足以下三个条件：

① 各地层应有各自的孔隙度—深度曲线，这主要是因为不同地层的岩性和结构存在差异。

② 用今天实测的孔隙度—深度曲线代替古孔隙度—深度曲线，但今测的孔隙度—深度曲线可能受到各种地质事件(如剥蚀、断层等)的影响，不能真实地反映正常压实下的孔隙度随深度的变化规律，因此必须消除这些影响，获得正常压实状况下的孔隙度—深度曲线。

③ 各地层应有三种岩性的孔隙度—深度曲线(砂岩、泥岩、灰岩)，其中砂岩假定为除泥岩和灰岩外的所有其它岩性的总和。

1. 骨架厚度公式和地层底界公式的提出

随着埋藏深度的增加，地层的上覆负载也增加，导致孔隙度变小，体积变小。我们假定，地层的横向在沉积过程中保持不变，地层被压实仅仅是纵向上变化。因此，地层体积的变小就归结为地层厚度的变小。另外，根据岩石骨架不可压缩的假设，地层的骨架厚度(孔隙度为零时的地层厚度)始终不变，除非发生剥蚀等地质事件。地层的骨架厚度始终小于实际的地层厚度。计算地层骨架厚度的公式为：

$$h_s = \int_{z_1}^{z_2} [1 - \varphi(z)] \mathrm{d}z \tag{12-30}$$

式中 h_s——地层的骨架厚度，m；

z_1——地层的顶界深度，m；

z_2——地层的底界深度，m；

$\varphi(z)$——地层的孔隙度—深度曲线函数，小数。

$\varphi(z)$的通式可表示为：

$$\varphi(z) = P_s\varphi_s(z) + P_m\varphi_m(z) + P_l\varphi_l(z) \tag{12-31}$$

式中 P_s——地层的砂岩含量，小数；

P_m——地层的泥岩含量，小数；

P_l——地层的灰岩含量，小数；

$\varphi_s(z)$—— 地层砂岩孔隙度—深度曲线函数；

$\varphi_m(z)$ —— 地层泥岩孔隙度—深度曲线函数；

$\varphi_l(z)$ —— 地层灰岩孔隙度—深度曲线函数。

在上式中应有：$P_s+P_m+P_l=1$。$\varphi(z)$表示为砂岩、泥岩、灰岩三种岩性孔—深曲线函数的加权平均值。一般，砂岩、泥岩、灰岩三种岩性的孔隙度—深度曲线函数形式如下：

砂岩：

$$\varphi_s(z) = \varphi_{0s}e^{-c_s z} \tag{12-32}$$

式中 φ_{0s}—— 砂岩的初始($z = 0$) 孔隙度，小数；

c_s—— 常数，m^{-1}。

泥岩：

$$\varphi_m(z) = \varphi_{0m}e^{-c_m z} \tag{12-33}$$

式中 φ_{0m}——泥岩的初始($z=0$)孔隙度，小数；

c_m——常数，m^{-1}。

灰岩：

$$\varphi_l(z) = \varphi_{0l}e^{c_l z} \tag{12-34}$$

式中 φ_{0l}——灰岩的初始($z=0$)孔隙度，小数；

c_l——常数，m^{-1}。

将式(12-32)、(12-33)、(12-34)代入式(12-31)得：

$$\varphi(z) = P_s\varphi_{0s}e^{-c_s z} + P_m\varphi_{0m}e^{-c_m z} + P_l\varphi_{0l}e^{-c_l z} \tag{12-35}$$

将式(12-35)代入式(12-30)得：

$$\begin{aligned} h_s &= \int_{z_1}^{z_2} [1 - P_s\varphi_{0s}e^{-c_s z} - P_m\varphi_{0m}e^{-c_m z} - P_l\varphi_{0l}e^{-c_l z}] \mathrm{d}z \\ &= \left[z + \frac{P_s}{c_s}\varphi_{0s}e^{-c_s z} + \frac{P_m}{c_m}\varphi_{0m}e^{-c_m z} + \frac{P_l}{c_l}\varphi_{0l}e^{-c_l z}\right]_{z_1}^{z_2} \\ &= z_2 - z_1 + \frac{P_s}{c_s}\varphi_{0s}(e^{-c_s z_2} - e^{-c_s z_1}) + \frac{P_m}{P_m}\varphi_{0m}(e^{-c_m z_2} - e^{-c_m z_1}) \\ &\quad + \frac{P_l}{c_l}\varphi_{0l}(e^{-c_l z_2} - e^{-c_l z_1}) \end{aligned}$$

即

$$h_s = z_2 - z_1 + \frac{P_s\varphi_{0s}}{c_s}(e^{-c_s z_2} - e^{-c_s z_1}) + \frac{P_m\varphi_{0m}}{c_m}(e^{-c_m z_2} - e^{-c_m z_1}) + \frac{P_l\varphi_{0l}}{c_l}(e^{-c_l z_2} - e^{-c_l z_1}) \tag{12-36}$$

式(12-36)是考虑三种岩性(砂岩、泥岩、灰岩)的地层骨架厚度公式。是回剥技术中重要的公式之一。如果地层没有发生剥蚀,则根据目前资料求出的各地层的骨架厚度在以前的各地质时刻保持不变。

将式(12-36)变换位置可得:

$$z_2 = h_s + z_1 - \frac{P_s\varphi_{0s}}{c_s}(e^{-c_sz_2} - e^{-c_sz_1}) - \frac{P_m\varphi_{0m}}{c_m}(e^{-c_mz_2} - e^{-c_mz_1}) - \frac{P_l\varphi_{0l}}{c_l}(e^{-c_lz_2} - e^{-c_lz_1}) \tag{12-37}$$

式(12-37)就是考虑三种岩性(砂岩、泥岩、灰岩)的地层底界公式,在骨架厚度和地层顶界已知的情况下,可由该式直接计算出地层底界埋深,是回剥技术中重要的公式之一。

2. 地层骨架厚度和地层底界公式的变化

① 当地层的孔隙度一深度曲线用同一个函数难以描述而需要多个函数描述时,对原骨架厚度公式必须修改。

例如:如图 12-6 所示,某井某地层的砂、泥、灰岩孔隙度一深度曲线相同,由三个函数描述:

$$\varphi(z)=\begin{cases}0.5e^{(-0.523\times10^{-3}z)} & 0\leqslant z\leqslant1900\\ 0.185 & 1900\leqslant z\leqslant2100\\ 0.41e^{(-0.379\times10^{-3}z)} & z\geqslant2100\end{cases}$$

这时,必须对骨架厚度公式分段积分。若所求地层的顶界和底界分别为 1500 米和 2600 米,则有:

$$h_s = \int_{1500}^{2600}(1-\varphi(z))\mathrm{d}z$$

$$= \int_{1500}^{1900}(1-0.5e^{-0.523\times10^{-3}z})\mathrm{d}z + \int_{2100}^{2600}(1-0.185)\mathrm{d}z + \int_{2100}^{2600}(1-0.41e^{-0.379\times10^{-3}z})\mathrm{d}z$$

此时,回剥中的两个重要公式式(12-36)和式(12-37)必须做相应的修改。这个推导是比较容易的。

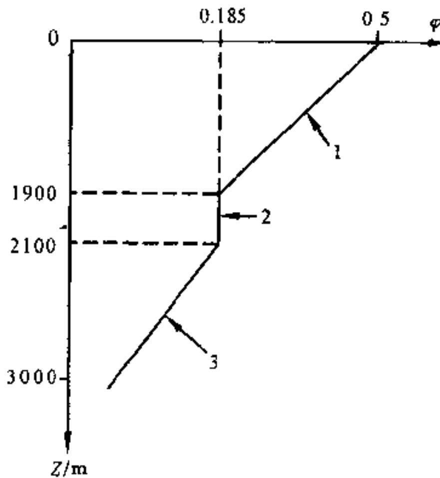


图 12-6 分段表达的孔隙度-深度曲线

② 当常数项 c_s、c_m、c_l 中有 0 出现时,必须修改骨架厚度公式和地层底界公式。

若 $c_m=0, c_s\neq0, c_l\neq0$,则:$\varphi_m(z)=\varphi_{0m}$,从而有:

$$h_s = \int_{z_1}^{z_2}[1 - P_s\varphi_{0s}e^{-c_sz} - P_m\varphi_{0m} - P_l\varphi_{0l}e^{-c_lz}]\mathrm{d}z$$

$$= z_2 - z_1 + \frac{P_s}{c_s}\varphi_{0s}(e^{-c_sz_2} - e^{-c_sz_1}) - (z_2-z_1)P_m\varphi_{0m} + \frac{P_l}{c_l}\varphi_{0l}(e^{-c_lz_2} - e^{-c_lz_1})$$

仅仅是其中第二项发生了变化,即:

当 $c_m=0$,以 $-(z_2-z_1)P_m\varphi_{0m}$ 取代 $\frac{P_m\varphi_{0m}}{c_m}(e^{-c_mz_2}-e^{-c_mz_1})$ 项。

依此类推得：

当 $c_s=0$，以 $-(z_2-z_1)P_s\varphi_{0s}$ 取代 $\frac{P_s\varphi_{0s}}{c_s}(e^{-c_s z_2}-e^{-c_s z_1})$ 项。

当 $c_l=0$，以 $-(z_2-z_1)P_l\varphi_{0l}$ 取代 $\frac{P_l\varphi_{0l}}{c_l}(e^{-c_l z_2}-e^{-c_l z_1})$ 项。

地层的底界公式也应做相应的改动。

3. 地层底界方程的求解

地层底界方程(12-37)中，若地层骨架厚度 h_s 和地层顶界 z_1 为已知，则方程形式可归结为：

$$z_2 = f(z_2)$$

上述方程左右两端含有未知数 z_2，对于这种形式的方程，可选用迭代法求解。迭代公式如下：

$$z_2^{(k)}=f(z_2^{(k-1)}) \qquad (k=1,2,3\cdots)$$

其中的迭代初值 $z_2^{(0)}$ 一般取该地层的顶界加上其骨架厚度。根据迭代公式可以得到：$z_2^{(1)}$，$z_2^{(2)}$，$z_2^{(3)}$，…，$z_2^{(k)}$，$z_2^{(k+1)}$，…，直至满足 $\left|z_2^{(k+1)}-z_2^{(k)}\right|<\varepsilon(10^{-5})$ 为止，取 $z_2=z_2^{(k+1)}$ 为所求地层的底界。

4. 单井逐层回剥过程(假设无剥蚀、无断层)

单井逐层回剥最终形成单井回剥柱状剖面图。假设目前盆地从上到下共有5套地层，代号为1、2、3、4、5，单井逐层回剥过程一般应经过以下几步：

① 确定有关参数。主要包括：各套地层的底界地质年龄(距今时间) t_1，t_2，t_3，t_4，t_5，各层的底界深度，各层泥岩、砂岩、灰岩的含量，各层的地层孔隙度—深度曲线(泥岩、砂岩、灰岩)。

② 确定所使用的地层骨架厚度公式和地层底界公式，求各地层的骨架厚度。

③ 取时刻 t_1，剥第一层。此时：

第二层的顶界＝0，按底界公式求出第二层的底界深度。

第三层的顶界深度＝第二层的底界深度，按底界公式求出第三层的底界深度。

第四层的顶界深度＝第三层的底界深度，按底界公式求出第四层的底界深度。

第五层的顶界深度＝第四层的底界深度，按底界公式求出第五层的底界深度。

④ 取时刻 t_2，剥第二层。此时：

第三层的顶界＝0，按底界公式求出第三层的底界深度。

第四层的顶界深度＝第三层的底界深度，按底界公式求出第四层的底界深度。

第五层的顶界深度＝第四层的底界深度，按底界公式求出第五层的底界深度。

⑤ 取时刻 t_3，剥第三层。此时：

第四层的顶界＝0，按底界公式求出第四层的底界深度。

第五层的顶界深度＝第四层的底界深度，按底界公式求出第五层的底界深度。

⑥ 取时刻 t_4，剥第四层。此时：

第五层的顶界＝0，按底界公式求出第五层的底界深度。

⑦ 取时刻 t_5，此时无沉积。

按上述过程，我们可以绘制单井回剥柱状剖面图(如图12-7)。

通过对盆地内每口人工井进行类似的回剥模拟，将所有的单井柱状剖面在平面上联系起来，即可达到研究全盆地沉积发育史的目的。可以绘制的图件包括盆地演化宝塔图、盆地发育

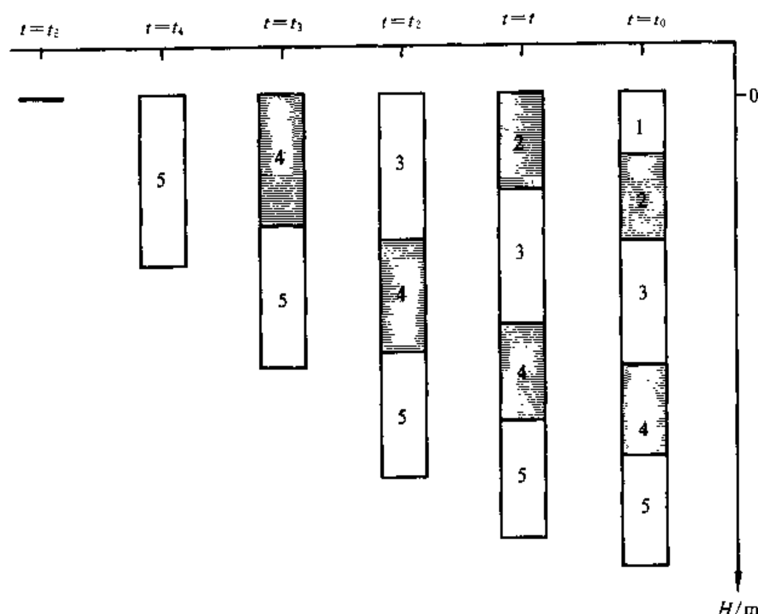


图 12-7　单井回剥柱状剖面图

剖面图、单井埋藏史曲线图等。

5. *回剥过程中的剥蚀事件*

上述回剥过程是在假设地层没有剥蚀的情况下进行的，如果地层存在剥蚀，则地层回剥柱状图中应增加剥蚀时间开始时的时间刻度及相应的地层柱。由于回剥过程中必须计算地层的骨架厚度，而计算地层的骨架厚度是按地层的顶界和底界所达到的历史最大深度进行的，且骨架厚度在地层埋深较浅时保持不变。因此考虑剥蚀事件时必须区分二种情况：一是在剥蚀事件发生后，目前地层的埋深仍是历史最大埋深，这时的剥蚀量相对较小。二是剥蚀事件导致目前地层的埋深不是历史最大埋深，这时的剥蚀量相对较大。

(1) 目前地层的埋深是历史最大埋深的情况

在这种情况下，可直接按现有的分层及其他资料计算各地层骨架厚度。假设某盆地有 n 组地层，其底界目前地质年龄由新到老分别为：t_1、t_2、…、t_n，根据目前各地层底界深度（从地表深度为 0 起算）计算出各地层目前的骨架厚度为：h_s^1、h_s^2、…、h_s^n，其中第 i 层存在剥蚀，所计算出该层的骨架厚度为残余骨架厚度，剥蚀厚度为 h_i，剥蚀开始的时间为 t_e，($t_{i-1}<t_e<t_i$)。为了反映地史过程中的剥蚀事件，回剥过程中的时间序列可考虑为：t_0、t_1、t_2、…、t_{i-1}、t_e、t_i、…、t_n，其中 t_0 表示目前时间。

按正常情况下的单井回剥过程可连续剥去 $i-1$ 个层，这时的时间刻度为 t_{i-1}，即第 i 个地层的剥蚀事件刚刚结束的时间、所求的是该层剥蚀后的残余部分的底界。此时应注意计算方法，第 i 层的底界是按顶界深度等于剥蚀厚度，骨架厚度等于残余骨架厚度，使用地层底界方程(12-37)计算得到未剥蚀时第 i 层的底界深度后再减去剥蚀厚度得到。这时计算出的就是第 i 层剥蚀后残余部分的底界深度。为什么取顶界深度等于剥蚀厚度呢？根据剥蚀过程中地层抬升，地层厚度保持不变的原理，剥蚀后残余地层厚度等于未剥蚀时这部分地层的厚度，而未剥蚀时这部分地层的顶界深度可认为等于剥蚀厚度。接下来取第 i 层的剥蚀开始时间 t_e，此时，第 i 层尚未开始剥蚀，第 i 个地层的顶界深度＝0，同样是考虑到地层抬升，厚度不变的原理，

底界深度可直接取为：

第 i 层的剥蚀厚度 h_i+t_{i-1}时刻该地层残余部分的底界深度

第 i 层底界深度求出后，可连续求出此时其它地层的底界深度，可得 $t=t_e$ 时的地层柱状剖面。随后依次取时间 t_i、…、t_n，按正常情况下的单井回剥过程剥去其它地层，直至回剥过程结束。有时，回剥过程中需要恢复被剥蚀地层在未被剥蚀时的原始骨架厚度，恢复的方法是在剥蚀后的残余骨架厚度（根据目前资料算出）的基础上增加被剥蚀部分的骨架厚度，被剥蚀部分的骨架厚度计算也使用地层骨架厚度公式(12-36)，此时必须注意被剥蚀部分的顶、底界应取所达到的历史最大深度。

如果有多个地层存在剥蚀，则应确定各剥蚀地层相应的剥蚀开始时间，并将这些时间加入到回剥过程的时间序列中去，然后按上述方法进行回剥。如果有地层连续剥蚀的情况，则必然存在被完全剥蚀的地层，这些地层目前的顶、底界深度取为相同，且骨架厚度为 0，在被完全剥蚀的地层开始剥蚀的时刻，该层的顶界深度为 0，底界深度等于剥蚀厚度，该层以下地层的底界深度应在上一时刻的基础上加上该地层的剥蚀厚度，而上一时刻各层底界的计算已考虑到剥蚀厚度，此时不能再用地层底界公式(12-37)求解各层的底界深度。这也是考虑到在剥蚀过程中，地层抬升，厚度不变的原理。

例如某井有六组地层，代号分别为：Q+N，E_d，E_s^1，E_s^2，E_s^3，E_s^4+K。其底界地质年龄分别为：22、30、32.5、35、37.5、54Ma，底界深度分别为：1000、1200、1600、2400、2400、3000 m，由式(12-36)计算出骨架厚度分别为：596.9、143.5、304.3、654.5、0、561.1 m，其中 E_d、E_s^3、E_s^4+K 三组地层存在剥蚀，剥蚀厚度分别为：100、500、40m、E_s^3 被完全剥蚀，且 E_s^3 剥蚀完后即开始剥蚀 E_s^4+K，剥蚀开始年代分别为：E_d～23.72Ma，E_s^3～36.25Ma，E_s^4+K～35.09 Ma。这六组地层的泥岩含量分别为：0.481、0.39、0.35、0.43、0.40、0.643，灰岩含量均为 0.1，砂岩含量分别为：0.419、0.51、0.55、0.47、0.50、0.257，这六组地层均采用相同的孔隙度一深度曲线。

砂岩孔隙度一深度曲线：

$$\varphi(z)=\begin{cases}0.5e^{(-0.523\times10^{-3}z)} & 0\leqslant z\leqslant 1900\\ 0.185 & 1900\leqslant z\leqslant 2100\\ 0.41e^{(-0.379\times10^{-3}z)} & z\geqslant 2100\end{cases}$$

泥岩孔隙度一深度曲线：

$$\varphi(z)=\begin{cases}0.5e^{(-0.603\times10^{-3}z)} & 0\leqslant z\leqslant 1900\\ 0.175 & 1900\leqslant z\leqslant 2100\\ 0.416e^{(-0.412\times10^{-3}z)} & z\geqslant 2\,100\end{cases}$$

灰岩孔隙度一深度曲线和泥岩相同。

整理后的时间序列为：0，22，23.72，30，32.5，35，35.09，36.25，37.5，54Ma。

利用上述参数进行地层回剥，回剥过程如下：

① 取 $t=0$(今天)。各层的底界深度分别为：1000，1200，1600，2400，2400，3000 m。

② 取 $t=22$Ma(Q+N 未沉积，E_d 剥蚀结束)，剥去 Q+N。E_d 顶界为 0，先求 E_d 残余部分的底界深度。按顶界(z_1)等于 100 m(剥蚀厚度)，残余骨架厚度等于 143.35 m，由地层底界公式(12-37)求出未剥蚀时 E_d 的底界深度为 366 m，由于此时 E_d 剥蚀结束，剥蚀厚度为 100 m，因此 E_d 残余部分的底界深度为 366 m－100 m＝266 m，进一步依次求出 E_s^1，E_s^2，E_s^3，E_s^4+K 的底界深度分别为：751，1 631，1 631，2257 m。

③ 取 t=23.72Ma (E_d 沉积结束,剥蚀欲开始),不剥去。E_d 底界深度为:100(剥蚀厚度)+266(上一时刻的底界深度)=366 m,将 100(剥蚀厚度)加到其它地层的底界深度,依次求出 E_s^1,E_s^2,E_s^3,E_s^4+K 的底界深度分别为:851,1731,1731,2357 m。

如果要恢复 E_d 的原始骨架厚度,先要求出被剥蚀部分的骨架厚度。可按 E_d 被剥蚀部分的顶、底界达到的最大深度 0m 和 100m,求出被剥蚀部分的骨架厚度为 49 m。

因此得到 E_d 未被剥蚀情况下的原始骨架厚度为:143.5 m+49 m=192.5 m。

④ 取 t=30Ma (E_s^1 沉积结束,E_d 未开始沉积),剥去 E_d。E_s^1 的顶界深度为 0m,骨架厚度为 304.3m,根据地层底界公式求出 E_s^1 底界深度为 553m,依次求出 E_s^2,E_s^3,E_s^4+K 的底界深度分别为:1486,1486,2125m。

⑤ 取 t=32.5Ma (E_s^2 沉积结束,E_s^1 未开始沉积),剥去 E_s^1。E_s^2 的顶界深度为 0m,骨架厚度为 654.5m,根据地层底界公式求出 E_s^2 界深度为 1080m,依次求出 E_s^3,E_s^4+K 的底界深度分别为:1756,1756m。

⑥ 取 t=35Ma (E_s^2 未开始沉积,E_s^4+K 剥蚀结束),剥去 E_s^2。E_s^4+K 的顶界为 0,求 E_s^4+K 残余部分的底界深度。按 E_s^4+K 顶界深度等于剥蚀厚度 540m(500m+40m),残余骨架厚度等于 516.1m,根据地层底界公式(12-37)求出 E_s^3 未剥蚀时 E_s^4+K 底界深度为 1296m。而 E_s^4-K 剥蚀结束后残余部分的底界深度为:1296m-540m(总剥蚀厚度)=756m。

⑦ 取 t=35.09Ma (E_s^3 剥蚀结束,E_s^4+K 欲开始剥蚀),不剥去。E_s^4+K 顶界深度为 0,底界深度为:40(剥蚀厚度)+756(上一时刻的底界深度)=796m。也可认为底界深度为:1296m-500m(E_s^3 剥蚀厚度)=796m。

⑧ 取 t=35.25Ma (E_s^3 沉积结束,剥蚀欲开始),增加 E_s^3。E_s^3 顶界深度为 0,底界深度为 500m(剥蚀厚度),E_s^4+K 的底界为:796+500=1296m。

⑨ 取 t=37.5Ma (Es_4+K 沉积结束,E_s^3 未沉积),剥去 E_s^3。E_s^4+K 顶界深度为 0,另外由于 E_s^4+K 增加了被剥蚀的部分,因此必须重新求解 E_s^4+K 的骨架厚度,即在原骨架厚度的基础上增加被剥蚀部分的骨架厚度。由于被剥蚀部分顶、底界达到的最大埋深是 500 和 540m,按骨架厚度公式(12-36)求出相应的骨架厚度为 24.1m,因此 E_s^4+K 未被剥蚀情况下的原始骨架厚度应为:516.1m+24.1m=540.2m。根据地层底界公式(12-37)求出 E_s^4+K 底界深度等于 924m。

⑩ 取 t=54Ma (无沉积),回剥结束。

地层回剥柱状图如图 12-8。

(2)目前地层的埋深不是历史最大埋深的情况

这种情况出现的原因可认为是剥蚀事件造成的,回剥的关键在于求各地层底界最大埋深和相应的骨架厚度。求解时必须注意以下二点原则:

① 地层的骨架厚度应按最大埋深时的顶、底界深度计算。

② 如果地层在某时期达到了最大埋深,则该地层在此后任一埋深更浅时刻的厚度保持不变,不能通过地层底界公式再次进行迭代计算。

如果目前的地层不是处于历史最大埋深,则最大埋深一般出现在剥蚀事件开始的地质年代。确定了相应的地质年代和总剥蚀厚度后,采用迭代法求解地层底界最大埋深和骨架厚度(可能是残余骨架厚度)。首先由目前地层的骨架厚度和剥蚀厚度求地层在最大埋深时刻(未开始剥蚀时刻)的底界,重新计算地层的骨架厚度(可能是残余骨架厚度),根据新的骨架厚度再求地层底界深度,然后再求骨架厚度和相应的地层底界,…,如此迭代计算,直至前后所求地层

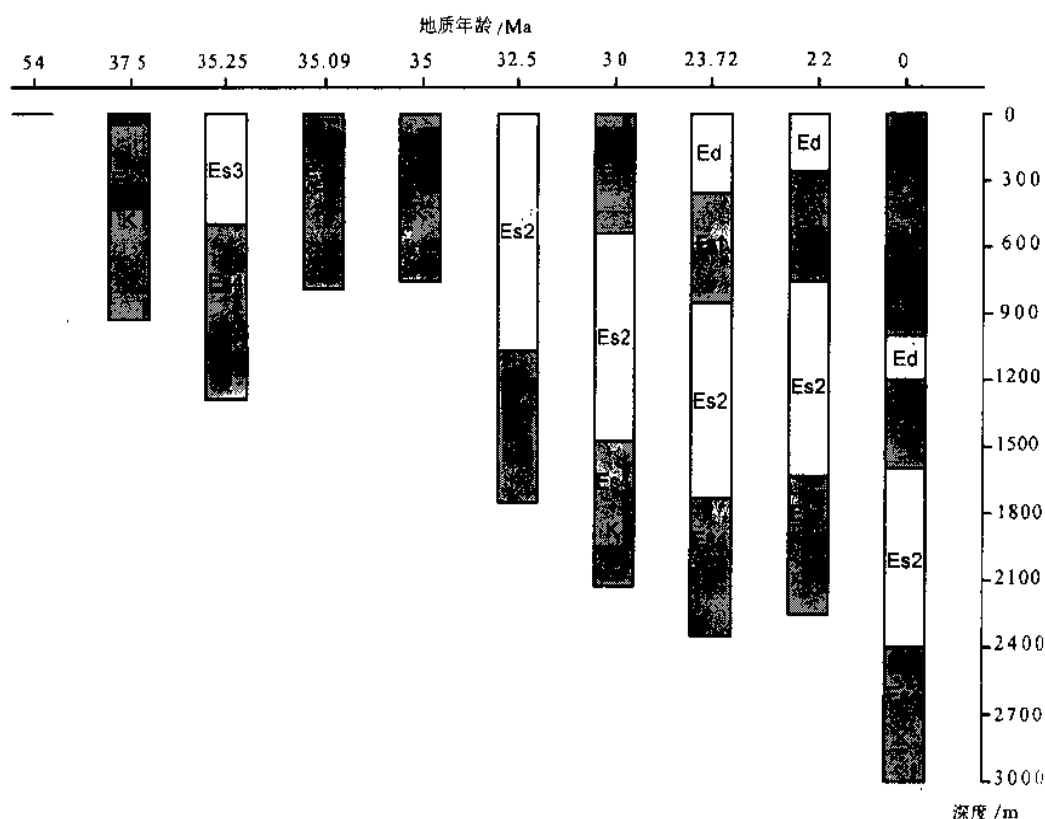


图 12-8　某模型井单井回剥柱状剖面图

底界值之差在给定的足够小的误差范围之内。最后一次求出的地层底界深度和骨架厚度就是所求值。

对求出了历史最大埋深时地层的骨架厚度后，具体回剥过程与前一种情况相同。

(3) 剥蚀时间的计算

一般情况下，各剥蚀事件的开始时间应作为已知参数给出，如果无法给出，可采用预测公式进行近似计算。

假设：被剥蚀地层的剥蚀开始年代为 t_e/Ma，剥蚀厚度为 h_e/m，顶界地质年龄为 t_i/Ma，底界的地质年龄为 t_{i+1}/Ma，被剥蚀地层 t_i时的残余厚度为 h_i/m，如果认为沉积速率等于剥蚀速率，则下式成立：

$$\frac{h_e + h_i}{t_{i+1} - t_e} = \frac{h_e}{t_e - t_i}$$

整理后可得：

$$t_e = \frac{h_e t_{i+1} + (h_i + h_e)t_i}{h_i + 2h_e} \tag{12-38}$$

上式即是预测剥蚀开始时间的计算公式，式中 $t_i < t_e < t_{i+1}$。上式是在沉积速率等于剥蚀速率的前提下推出的，如果这两种速率不相等，也不难推出类似的计算公式。

对于连续剥蚀的地层，除使用式(12-38)进行计算外，还要使用另外的计算公式。设有 m

个连续剥蚀的地层，它们的剥蚀量分别为：$h_{e1},h_{e2},\cdots,h_{em}$，剥蚀开始年代分别为：$t_{e1},t_{e2},\cdots,t_{em}$，且有：$t_{e1}>t_{e2}>\cdots>t_{em}$。令 $h_e=h_{e1}+h_{e2}+\cdots+h_{em}$，将 h_e 代入式(12-38)可计算出 t_{e1}，即 $t_{e1}=t_e$。$t_{e2},\cdots,t_{em}$可由下式算出：

$$t_{ek}=t_e-\frac{t_e-t_i}{h_e}\sum_{l=1}^{k-1}h_{el}\qquad(k=2,3,\cdots,m)\tag{12-39}$$

不难看出，上式是在各地层的剥蚀速率相同的前提下，按剥蚀厚度比例推导得出的。

6. 超压技术(欠压实带适用)简介

前面我们所介绍的逐层回剥技术，其适用范围仅仅是正常压实带，在欠压实带使用逐层回剥技术可能影响到地层的厚度史恢复的精度，这主要是由于孔－深曲线的不准确性所致。本节所介绍的超压技术，正是修正或减小厚度史误差的一种方法。也就是说，是修正用回剥技术求出的地层古厚度史的一种方法。

超压技术的思路：求出欠压实地层(主要是生油层)的古超压史和古厚度史，并用新的古厚度史代替用回剥技术求出的古厚度史，达到修正的目的。

超压技术中包括两个主要方程：

① 古超压方程。其作用是求出从古到今的生油层古超压史(正演)。

② 古厚度方程。其作用是求出从古到今的生油层古厚度史(正演)。

(1) 古超压方程和古厚度方程

古超压方程：$$\frac{\partial P_a}{\partial t}=G_\sigma\frac{\partial H}{\partial t}+\frac{G_\sigma}{\bar{c}}\cdot\frac{1}{h_s}\cdot\frac{(1-\varphi)^2}{\varphi_1}\cdot\frac{\partial h}{\partial t}\tag{12-40}$$

或 $$\frac{\partial P_a}{\partial t}=G_\sigma\frac{\partial H}{\partial t}+\frac{G_\sigma}{\bar{c}}\cdot\frac{1}{h_s}\cdot\frac{(1-\varphi)^2}{\varphi_1}\cdot\left[-\left(\frac{K_1}{\mu_1}+\frac{K_2}{\mu_2}\right)\frac{P_a}{0.5h}\right]\tag{12-41}$$

$$G_\sigma=(\bar{\rho}_s(1-\bar{\varphi})+\bar{\rho}_f\bar{\varphi}-\bar{\rho}_f)g$$

$$\bar{c}=P_s\bar{c}_s+P_m\bar{c}_m+P_l\bar{c}_l$$

古厚度方程：$$\frac{\partial h}{\partial t}=-\left[\frac{K_1}{\mu_1}+\frac{K_2}{\mu_2}\right]\frac{P_a}{0.5h}\tag{12-42}$$

式中 P_a——生油层的孔隙流体超压(中点)；

H——上覆沉积总厚度；

h_s——生油岩骨架厚度；

h——生油层厚度；

φ——生油层中点的孔隙度；

φ_1——生油层顶界的孔隙度；

$\bar{\rho}_s$——上覆沉积物骨架的平均密度；

$\bar{\rho}_f$——上覆沉积物中流体的平均密度；

ρ_f——生油层中点的孔隙流体密度；

$\bar{\varphi}$——上覆沉积物的平均孔隙度；

$\bar{c}_s,\bar{c}_m,\bar{c}_l$——上覆沉积中三类岩石孔－深曲线中的 c_s,c_m,c_l的平均值；

P_s,P_m,P_l——上覆沉积中三类岩石的平均含量；

g——重力加速度；

K_1、K_2——生油层顶、底界的渗透率；

μ_1、μ_2——生油层顶、底界的流体粘度。

(2) 古超压方程和古厚度方程的求解

①古超压方程的求解

对古超压方程(12-41)按时间差分可得：

$$P_{ak+1}-P_{ak}=G_{\sigma k+\frac{1}{2}}(H_{k+1}+H_k)+\frac{G_{\sigma k+\frac{1}{2}}}{\bar{c}_{k+\frac{1}{2}}}\cdot\frac{1}{h_s}\cdot\frac{(1-\varphi_{k+\frac{1}{2}})^2}{\varphi_{k+\frac{1}{2}}}\left[\left(\frac{K_1}{\mu_1}+\frac{K_2}{\mu_2}\right)\frac{1}{0.5h}\right]_{k+1}P_{ak+1}\cdot\Delta t \tag{12-43}$$

$k=1,2,\cdots\cdots$,目前(时间)

$k+\frac{1}{2}$为参数在 k 和 $k+1$ 时的平均值。在厘米·克·秒制的量纲下,再考虑转换系数,由式(12-43)可推导出 $k+1$ 时古超压的计算公式：

$$P_{ak+1}=\frac{P_{ak}+G_{\sigma k+\frac{1}{2}}(H_{k+1}-H_k)/1.01\times10^6}{1-\frac{G_{\sigma k+\frac{1}{2}}}{c_{k+\frac{1}{2}}}\cdot\frac{1}{h_s}\cdot\frac{(1-\varphi_{k+\frac{1}{2}})^2}{\varphi_{k+\frac{1}{2}}}\cdot\left[-\left(\frac{K_1}{\mu_1}+\frac{K_2}{\mu_2}\right)\frac{1}{0.5h}\right]_{k+1}\cdot\Delta t/1.01\times10^6} \tag{12-44}$$

式中　P_{ak+1}——生油层中点 $k+1$ 时的古超压,atm；

H——上覆沉积物的厚度,cm；

h_s——生油层骨架厚度,m；

φ——生油层中点的孔隙度,小数；

φ_1——生油层顶界的孔隙度,小数；

h——生油层厚度,m；

Δt——k 至 $k+1$ 的时间步长,s；

K_1——生油层顶界的渗透率,D(达西)；

K_2——生油层底界的渗透率,D(达西)；

μ_1——生油层顶界的流体沾度,cp(厘泊)；

μ_2——生油层底界的流体沾度,cp(厘泊)；

G_σ——岩石骨架有效应力的梯度,dyn /cm^2·cm；

1.01×10^6——转换系数,1atm $=1.01\times10^6$ dyn/cm^2；

上式中除 $K_1,K_2,\mu_1,\mu_2,G_\sigma$外均可由地史模拟计算得出。计算 $K_1,K_2,\mu_1,\mu_2,G_\sigma$采用以下公式。

1° 计算渗透率 K

$$K=\begin{cases}\dfrac{0.2\varphi^3}{S_a^2(1-\varphi)^2} & \varphi\geqslant0.1\\[2ex]\dfrac{20\varphi^5}{S_a^2(1-\varphi)^2} & \varphi<0.1\end{cases} \tag{12-45}$$

式中　K——渗透率,m^2,1D$=9.87\times10^{-13}$m^2；

φ——孔隙度,小数；

S_a——骨架的颗粒比面,m^2/m^3。

上式称为柯西—卡曼公式(Kozeny—Carman)。若考虑 K 的单位为达西,则上式又可以变为：

$$K=\begin{cases}\dfrac{0.2\varphi^3}{9.87\times10^{-13}S_a^2(1-\varphi)^2} & \varphi\geqslant0.1\\[2ex]\dfrac{20\varphi^5}{9.87\times10^{-13}S_a^2(1-\varphi)^2} & \varphi<0.1\end{cases}\qquad(12\text{-}46)$$

S_a的值的取值办法是：根据实测的多个样品的 K 和 φ，代入式(12-45)或式(12-46)，得到多个各样品的 S_a 值，取其平均值后做为使用的 S_a。鉴于地层岩性物性的非均质性，可考虑分小层确定 S_a。如某凹陷 E_3^3 的 S_a 在三个小层的值分别为：109799、109448、74883 m²/m³。

2° 计算流体粘度 μ

$$\mu=(5.3+3.8AT-0.26AT^3)^{-1}\qquad(12\text{-}47)$$

$$AT=(T-150)/100$$

式中 μ——流体粘度，cp；

T——古地温，℃，由热史模型确定。

3° 计算骨架有效应力 σ 的梯度 G_σ

$$G_\sigma=[\bar{\rho}_s(1-\bar{\varphi})+\bar{\rho}_f\bar{\varphi}-\rho_f]g\qquad(12\text{-}48)$$

式中 $\bar{\rho}_s$——上覆沉积物骨架的平均密度，g/cm³，可取为(2.55+2.61)/2；

$\bar{\varphi}$——上覆沉积物的平均孔隙度，小数；

ρ_f——生油层中点的孔隙流体平均密度，g/cm³；

g——重力加速度，cm/s²，可取为 981。

$\bar{\rho}_f$——上覆沉积物的孔隙流体的平均密度，g/cm³，$\bar{\rho}_f=\rho_{f0}[1-\beta_T(\bar{T}-T_s)]$；

ρ_{f0}——地表孔隙流体密度 g/cm³，可取 1.004；

β_T——流体热膨胀系数，℃⁻¹，可取 5×10^{-4}；

T_s——地表温度，℃；

$\bar{T}$——上覆沉积物的平均温度，℃，可取：$\bar{T}=(T_s+T)/2$

T——该层顶界温度，由热史模型确定。

由超压方程即可求得当 $k=1,2,\cdots$目前所有时刻的地层古超压。

② 古厚度方程的求解

对古厚度方程(12-42)按时间差分：

$$h_{k+1}=h_k+\left\{-\left[\frac{K_1}{\mu_1}+\frac{K_2}{\mu_2}\right]\frac{P_a}{0.5h}\right\}_{k+1}\Delta t=h_k+P_{ak+1}\left\{-\left[\frac{K_1}{\mu_1}+\frac{K_2}{\mu_2}\right]\frac{1}{0.5h}\right\}_{k+1}\Delta t\qquad(12\text{-}49)$$

$$k=1,2,\cdots,\text{目前(时间)}$$

式中 k——时间，$k=1$ 为生油层完成沉积的时间；

h——生油层的古厚度，cm；

其余符号和式(12-44)相同。

式(12-49)中 P_{ak+1}已由古超压方程求得，其余各参数也已求得，取定生油层的原始沉积厚度 h_1 后，即可求得：$h_2,h_3,\cdots h_{目前}$。从古超压方程和古厚度方程的求解可知，求解过程必须建立在回剥地史和温度史已求出的基础之上。

③ 古超压史和古厚度史的计算(生油层)

取计算的开始时间 $k=1$(生油层完成沉积、上覆地层开始沉积的时间)，此时：

$P_{a1}=0$

h_1= 回剥技术算出的生油层厚度(原始沉积厚度)

由古超压方程和古厚度方程计算生油层的古厚度史为:$h^1, h^2, \cdots h^*$。所计算的生油层古厚度史的目前厚度 h^* 可能不等于该层的实际厚度 h,这主要是因为骨架厚度公式受欠压实的影响导致计算结果不准确,因此必须修正生油层的骨架厚度,使得下式成立:

$$|h^* - h| < \varepsilon \tag{12-50}$$

式中 h^*——由超压技术得到的生油层古厚度史的目前厚度,cm;

h——生油层今天的实际厚度实测,cm;

ε——允许的误差,可取 0.1 cm。

骨架厚度 h_s 的迭代修正公式为:

$$h_s^{(n+1)} = h_s^{(n)} + (h - h^*)(1 - \varphi)/100 \tag{12-51}$$

式中 $h_s^{(n+1)}$——修正后的生油层骨架厚度,m;

$h_s^{(n)}$——修正前的生油层骨架厚度,m;

φ——生油层中心点今天的孔隙度,小数;

n——迭代号,$n=1,2,\cdots$。

取 $h_s^{(1)}$=由回剥技术算出的生油层骨架厚度 h_s,由古超压方程和古厚度方程可计算出该层的古超压史和古厚度史,如果满足式(12-50),这时算出的古超压史和古厚度史被认为是合乎要求的,如果不满足式(12-50),需要按式(12-51)修正骨架厚度公式,并重新计算古超压史和古厚度史,接着再判断是否满足式(12-50),如此迭代循环,直至修正的骨架厚度 h_s 使式(12-50)满足为止。对盆地内每套生油层(包括其他有欠压实的层)均用修正的生油层古厚度史代替回剥技术所得的生油层的古厚度史,并考虑其余各层顶界和底界的变化,即可达到修正地史的目的。

以上的过程是针对一个单井的生油层,如果对整个模拟地区中的所有人工井点均重复上述过程,就可获得该区各生油层的古超压史和古压力史,在平面上将其联系起来可获得古超压史和古压力史的各种图件,如超压平面等值线图,在该图的基础上还可绘制古超压平面流线图(在等值图上绘制从高线值指向低线值的箭头),该图指示了生油层的平面排烃方向。

7. *回剥与超压相结合的模拟过程(单井过程)*

由前所述,回剥技术适合于正常压实带,而超压技术适合于欠压实带,这两种技术相结合即可满足所有地层的古厚度史和古压力史求解。在单井模拟过程中,采用回剥与超压相结合的方法,过程如下:

① 用回剥技术计算地史(生油层厚度史可能有误差)。

② 用热史模型计算出该井的古地温史(为超压技术提供参数)。

③ 一个生油层古超压史和古厚度史的重新计算。

④ 地史修正(修正回剥所得地史);以该生油层的顶界埋深为基础,以修正后的古厚度史为尺度,可得该生油层的底界埋藏史,该生油层以下各层的底界埋藏史也应随之修正。

⑤ 对所有生油层从浅到深重复①~④步,最终获得经修正的单井地史。

修正后的地史考虑了正常压实和欠压实两种地质现象。

8. *关于埋藏点的加密*

无论是回剥技术还是超压技术,如果盆地内地层厚度过大,势必会影响到所计算地史的精度。为此,我们需要将过厚的地层划分为许多小段(相应的时间间隔也划分为若干小段),即埋藏点的加密。实际上,对盆地内的所有地层,均在加密后进行模拟。加密的一般方法是:取定一

个时间步长/Ma，将所有地层按时间步长划分若干小段，计算各小层的分层深度，形成新的分层资料。根据新的分层资料进行单井模拟。

关于时间步长的取定，中国东部盆地（埋藏时间较短）取 1～5 百万年，中国西部盆地（埋藏时间较长）取 5～10 百万年（据石广仁等）。

§2 热史模型

热史模型的功能是描述和重建含油气盆地的古热流史和古地温史。在盆地模拟系统中，热史模型作为一个关键模型，其作用在于为生烃史、排烃史和运移聚集史模拟提供温度场。其实，就热史模拟本身来说，也具有很大的地质意义。地热在沉积物的成岩、演化过程中起着重要的作用，各种岩石化学变化和矿物转化都以温度为重要条件。不仅如此，温度对于油气的保存和破坏也是具有普遍意义的控制条件。由此可见恢复盆地古热流史和古地温史在石油地质研究中的重要性。

地球内部的热传播方式包括热传导、热对流和辐射。一般是以热传导为主，热对流次之，热辐射再次之。热传播最主要的热源是来自地球软流圈的热流，称大地热流，这是普遍存在的热源。此外，还可能有局部的热源，如岩浆侵入带、断裂活动带、放射性元素富集区等。本节热史模型中所考虑的热源是大地热流。

热史模拟目前有两种方法，一是地球热力学方法，二是地球热力学和地球化学相结合的方法（结合法）。地球热力学方法适用于含油气盆地勘探早期。随着勘探程度的进一步深入和地球化学资料的逐渐丰富，地球热力学和地球化学相结合的方法在模拟精度方面更胜于单纯的地球热力学法。单纯的地球热力学法的不足之处是缺乏对古热流史的定量模拟，通常是在对盆地地热研究的基础上通过人为或类比等方式给出盆地各地质时期的古大地热流值，然后结合地球热力学温度史模型求出盆地各地层的古地温史，如果模拟出的温度史的今天值与实测资料不符，要返回去调整古大地热流值或其他热学参数，直至温度检验满足为止。地球热力学和地球化学相结合的方法比单纯的地球热力学法更进一步的地方在于能利用地球化学资料，在目前大地热流值的基础上确定盆地各地质时期的古大地热流值，即古热流史模拟，而古热流史是直接影响温度史模型最重要的参数之一。由于结合法是从今天的热流值入手，故温度史模拟的结果和今天的实测温度吻合情况较好。

从热史模型的功能上看，它应该包括二方面的内容：模拟研究盆地的大地热流史和研究盆地内各地层的古地温史。因此，热史模型一般由下面二个子模型构成：

① 温度史模型。

② 古热流史模型。

由于在盆地模拟中关于生油岩中有机质成熟度史的模拟和热史模拟是同步进行的，所以也有的盆地模拟系统把成熟度史模型归到热史模型之中，这样热史模型又增加了一个子模型，即成熟度史模型。其实，模型归属哪一类不是问题的关键，关键是如何建立正确的模拟模型。下面我们分别就地球热力学法以及地球热力学和地球化学相结合的方法介绍温度史模型和古热流史模型。

一、地球热力学法

主要介绍地球热力学方法下的温度史模型和古热流确定方法。

（一）温度史模型

1. 能量守恒热流方程(Stallman. R. W. 1967)

(1) 数学模型

描述能量守恒定律的热流方程控制着能量的传递，Stallman. R. W 据能量守恒定律推导了由热的传导和热对流两者同时发生的热流方程：

$$\nabla(K_s\nabla T_m)-c_w\rho_w\nabla(\vec{V}T_m)+Q=c_s\rho_s\frac{\partial T_m}{\partial t} \tag{12-52}$$

（热传导）　　（热对流）　　（热源）（热的净聚集）

式中 T_m—— 温度，℃；

K_s—— 沉积物的热导率，μcal/ (cm・s・℃)；

$\vec{V}$——流体流动速度，cm/s；

c_w、c_s—— 流体和沉积物的比热，μcal/ (g・℃)；

ρ_w、ρ_s——流体和沉积物的密度，g/cm³；

Q ——热源或热汇项(大地热流)，HFU=μcal/ (cm²・s)；

t——以地层从地表开始沉积时间为 0 起算，直至今天的时间坐标。

该方程的特点是即考虑了热的传导又考虑了热的对流。

在上述热流方程中，K_s、c_s、ρ_w、ρ_s需采用公式进行计算：

1° 计算 K_s

$$K_s=K_r\left(\frac{K_w}{K_r}\right)^{\varphi} \tag{12-53}$$

式中 K_r——岩石骨架的热率，μcal/ (cm・s・℃)，可取 5.1×103～6.3×10³；

K_w——孔隙流体的热导率，μcal/ (cm・s・℃)，可取 1.348×10³；

φ——沉积物的孔隙度，小数。

用上式可求任一深度下的沉积物热导率。

2° 计算 c_s

$$c_s=(1-\varphi)c_r[1+\Omega_r(T-T_0)]+\varphi c_w[1+\Omega_w(T-T_0)] \tag{12-54}$$

式中 c_s—— 沉积物比热，cal/ (g・℃)；

c_r—— 岩石骨架比热，cal/ (g・℃)，0.219～0.214；

c_w—— 孔隙流体比热，cal/ (g・℃)，1.008；

Ω_r，Ω_w—— 分别描述岩石骨架和流体的比热随温度的变化常数，可取 $\Omega_r=0.76\times10^{-3}$，$\Omega_w=0.219\times10^{-3}$；

T_0—— 地表平均温度，℃；

T—— 地层温度，℃；

φ——沉积物孔隙度，小数。

由上式即可求得任一深度下的沉积物的比热。

3° 计算 ρ_w

$$\rho_w=\rho_{w0}[1+\beta_T(T-T_0)] \tag{12-55}$$

式中 ρ_w—— 地层温度为 T 时，孔隙流体的密度，g/cm³；

ρ_{w0}——地表孔隙流体密度，g/cm³，可取 1.004；

β_T——流体受热膨涨系数，可取 0.5×10^{-3}；

T_0——地表温度，℃；

T—— 地层温度，C。

通过上式可求任何温度下的孔隙流体密度。

4° 计算 ρ_s

$$\rho_s = \varphi\rho_w + (1-\varphi)\rho_r \tag{12-56}$$

式中 ρ_s—— 沉积物密度，g/cm³；

φ——沉积物孔隙度，小数；

ρ_w—— 孔隙流体密度，g/cm³，可取 1.004；

ρ_r—— 岩面骨架密度，g/cm³，可取 2.55～2.61。

通过上式可求任何孔隙度和岩性沉积物的密度。

确定了所需的各项参数之后，即可采用有限差分方法求解热流方程(12-52)，最终求出盆地各地层的古温度史。

也有的观点认为，热的对流与热的传导相比，已小到可以忽略不计的地步，因此可将热流方程(12-52)简化为：

$$\nabla(K_s \nabla T_m) + Q = c_s\rho_s \frac{\partial T_m}{\partial t} \tag{12-57}$$

该式是关于热传导的方程，仍采用有限差分方法求解。在静水条件下的正常压实带，由于 $\vec{V} \approx 0$，从热流方程(12-52)中可知，上式可认为是式(12-52)在静水条件下的正常压实带的特例。模拟运算证明，在许多地区式由(12-52)和式(12-57)所得的温度史没有显著的差别。

(2) Stallman 热流方程方程的数值解法

在一维情况(单井模拟)下，Stallman 热流方程方程形式为：

$$\frac{\partial}{\partial z}\left(K_s \frac{\partial T_m}{\partial z}\right) - c_w\rho_w \frac{\partial}{\partial z}(V_z T_m) + Q = c_s\rho_s \frac{\partial T_m}{\partial t} \tag{12-58}$$

我们将盆地的下边界视为封闭边界，盆地上边界湖底的温度可取该区当时年平均地表温度，因此上述偏微分方程(12-58)的边界条件可考虑为：

$$\frac{\partial T_m}{\partial z} = 0 \qquad (z=0)$$

$$T_m = \text{年均地表温度} \qquad (z=M)$$

一维热流方程(12-58)的定解问题归结为：

$$\begin{cases} \dfrac{\partial}{\partial z}\left(K_s \dfrac{\partial T_m}{\partial z}\right) - c_w\rho_w \dfrac{\partial}{\partial z}(V_z T_m) + Q = c_s\rho_s \dfrac{\partial T_m}{\partial t} \\ \dfrac{\partial T_m}{\partial z} = 0 & z=0 \\ T_m = \text{年均地表温度} & z=M \end{cases} \tag{12-59}$$

采用有限差分方法求解上述定解问题，差分格式为：

$$\Delta K_s \Delta T_{m2} - (c_w\rho_w)^i V^{i+\frac{1}{2}}\left(\frac{T_{m2}^{i+1} + T_{m2}^i}{2}\right) + (c_w\rho_w)^{i-1} V^{i-\frac{1}{2}}\left(\frac{T_{m2}^i + T_{m2}^{i-1}}{2}\right) + Q^i$$
$$= (c_s\rho_s)^i \Delta Z_k^i (T_{m2}^i - T_{m1}^i) \tag{12-60}$$

式中
$$\Delta K_s \Delta T_{m2} = K_s^{i+\frac{1}{2}}(T_{m2}^{i+1} - T_{m2}^i) + K_s^{i-\frac{1}{2}}(T_{m2}^{i-1} - T_{m2}^i)$$

$$K_s^{i+\frac{1}{2}} = \frac{4K_s^{i+1}K_s^i \Delta t_m}{(K_s^{i+1} + K_s^i)(\Delta Z_k^{i+1} + \Delta Z_k^i)}$$

$$K_s^{i-\frac{1}{2}}=\frac{4K_s^iK_s^{i-1}\Delta t_m}{(K_s^i+K_s^{i-1})(\Delta Z_k^i+\Delta Z_k^{i-1})}$$

式中 i——空间网格节点的位置；

T_{m2}——本时间阶段温度值；

T_{m1}——上时间阶段温度值；

Δt_m——时间步长；

$\Delta Z^i=Z^i-Z^{i-1}$(空间步长)；

$V^{i+\frac{1}{2}}$——Δt_m 内通过单元上界面的流体体积，通过计算压实量获得；(见图 12-9)

$V^{i-\frac{1}{2}}$——Δt_m 内通过单元下界面的流体体积(见图 12 9)。

热流方程的差分格式可以整理为关于 T_m 的三对角线性方程组，采用追赶法求解，求解过程和超压方程完全一样，大家如果有兴趣的话可以自己推导，最终结果是时刻 t 时垂向各节点上的温度值：T_m^1、T_m^2、T_m^n… 。

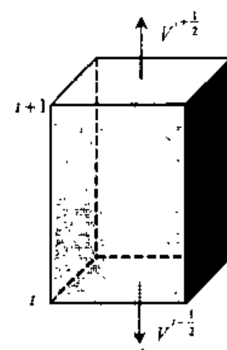


图 12-9 $V^{i+\frac{1}{2}}$和 $V^{i-\frac{1}{2}}$

2. 热传导温度史模型

与热的传导相比，热的对流对地层温度的影响要小得多。如果仅考虑热的传导，则我们可推导出下列的温度史模型。

(1) 数学模型

$$T(z,t)=T_0(t)+Q(t)\int_0^z\frac{1}{K(z,t)}\mathrm{d}z \qquad (12\text{-}61)$$

式中 t——地质时刻，百万年；

z——埋深，cm；

$T(z,t)$——时刻 t 埋深为 z 处的地层温度，℃；

$T_0(t)$——t 时刻地表温度，℃；

$Q(t)$——t 时刻的大地热流值，HFU=μcal/(cm²·s)；

$K(z,t)$——t 时刻深度 z 处的岩石热导率，μcal/(cm·s·℃)。

上述数学模型被许多盆地模拟系统所采用，其特点是只考虑了热的传导，并且计算简单快速，模拟效果良好。

(2) 数学模型推导

地热学中一般认为，来自于地下的热流垂直指向地面，并以辐射能的形式散失于空间。岩石的热导率在层内具各向同性。当热流速率达到平衡后，通过任何深度的热流值都是相同的。由于热流速率是地质时间 t 的函数，因此可认为大地热流也是时间 t 的函数。设地质时刻 t 的大地热流值为 $Q(t)$，则由地热学原理可知：

$$Q(t)=K\cdot\frac{\mathrm{d}T}{\mathrm{d}z} \qquad (12\text{-}62)$$

式中 K——岩石热导率；

T——温度；

z——深度；

$\frac{\mathrm{d}T}{\mathrm{d}z}$——地温梯度。

将上式变化为：

$$dT=Q(t)\cdot\frac{1}{K}dz$$

对上式两端取定积分：

$$\int_{\Omega_T}dT = Q(t)\int_{\Omega_z}\frac{1}{K}dz \tag{12-63}$$

式中 Ω_T——温度 T 的积分区间；

Ω_z——深度 z 的积分区间。

由于沉积物的岩性在垂向上是不断变化的，因此岩石的热导率也是不断变化的，而沉积物的厚度是地质时刻 t 的函数，因此可以认为，沉积物的热导率是埋深和地质时刻 t 的函数：

$$K=K(z,t)$$

设 t 时刻地表温度为 $T_0(t)$，埋深为 z 处地层温度为 $T(z,t)$。则当积分区间 $\Omega_z=[0,z]$ 时，可取定 $\Omega_T=[T_0(t),T(z,t)]$，从而式(12-63)可转化为：

$$\int_{T_0(t)}^{T(z,t)}dT = Q(t)\int_0^z\frac{1}{K(z,t)}dz$$

即
$$T(z,t)-T_0(t) = Q(t)\int_0^z\frac{1}{K(z,t)}dz$$

即
$$T(z,t) = T_0(t)+Q(t)\int_0^z\frac{1}{K(z,t)}dz$$

正是式(12-61)，推导结束。

上述温度史数学模型是一维的，也可以推广为三维情况下的温度史模型：

$$T(x,y,z,t) = T_0(x,y,t)+Q(x,y,t)\int_0^z\frac{1}{K(x,y,z,t)}dz \tag{12-64}$$

(3) 热传导温度史方程数值解法

采用数值积分的方法求解温度史方程(12-61)，求解的关键是对方程右端定积分的求解。假设在地质时刻 t，盆地最早沉积的地层的底界埋深为 z，使用正演法地史模拟的垂向节点划分方法，结合地史模拟求得的各节点的埋深 $z_i(i=1,2,\cdots,m+1)$，其中 $z_1=z$，$z_{m+1}=0$。由此可以将深度区间[0，z]划分 m 为个小区间[z_i，z_{i+1}]，$i=1,2,\cdots,m$，当垂向步长不是很大时，可考虑采用简单的数值积分方法，如梯形公式法和矩形法等。

1° 梯形公式法

$$\int_0^z\frac{1}{K(z,t)}dz = \sum_{i=1}^m\frac{1}{2}\left(\frac{1}{K(z_{i+1},t)}+\frac{1}{K(z_i,t)}\right)(z_{i+1}-z_i)$$
$$=0.5\sum_{i=1}^m\frac{K(z_{i+1},t)+K(z_i,t)}{K(z_{i+1},t)\cdot K(z_i,t)}(z_{i+1}-z_i)$$

故有
$$T(z,t) = T_0(t)+0.5\cdot Q(t)\cdot\sum_{i=1}^m\frac{K(z_{i-1},t)+K(z_i,t)}{K(z_{i+1},t)\cdot K(z_i,t)}(z_{i+1}-z_i)$$

式中 $K(z_i,t)$——时刻 t 埋深为 z_i 处的沉积物热导率，μcal/(cm·s·℃)。

热导率 $K(z_i,t)$采用下式计算：

$$K(z_i,t) = \left(\frac{K_w(t)}{K_r(z_i,t)}\right)^{\varphi(z_i,t)} \tag{12-65}$$

式中　$K_w(t)$——t 时刻孔隙流体的热导率，可取 $1.348\times10^3\mu cal/(cm\cdot s\cdot ℃)$；

$K(z_i,t)$——t 时刻深 z_i 处岩石骨架热导率，$5.1\times10^3\sim6.3\times10^3\mu cal/(cm\cdot s\cdot ℃)$；

$\varphi(z_i,t)$——t 时刻深 z_i 处沉积物孔隙度，由地史模型给出；

$T_0(t)$——年平均地表温度，可取目前平均地表温度；

$Q(t)$——t 时刻大地热流值，由古热流史模型求得。

2° 矩形公式法

$$\int_{K(z,t)}^{1}\frac{1}{K(z,t)}\mathrm{d}z=\sum_{i=1}^{m}(z_{i+1}-z_i)\cdot\frac{1}{K(z_i,t)}=\sum_{i=1}^{m}\frac{z_{i+1}-z_i}{K(z_i,t)}$$

故有

$$T(z,t)=T_0(t)+Q(t)\cdot\sum_{i=i}^{m}\frac{z_{i+1}-z_i}{K(z_i,t)}$$

式中　$K(z_i,t)$——t 时刻深 z_i 处的沉积物热导率，$\mu cal/(cm\cdot s\cdot ℃)$，由式(12-65)计算。

（二）古大地热流值的确定

在单纯的地球热力学方法中，古大地热流史的模拟是薄弱环节，没有有效的方法研究古热流史模型。一般假设大地热流在一定时期内是恒定的，盆地发育过程中的许多地质阶段（时期），我们对古热流史模型所做的工作就是确定这些时期（各地层开始沉积到结束沉积）的大地热流。

确定古大地热流值最简单的办法就是对它直接给定，基础是对该地区地热问题的详细研究。例如目前有：$Q_0=1.8$HFU，若认为该区的大地古热流呈衰减趋势，则古大地热流值按地质时期由新到老可以取定为 1.7、1.6、1.55、1.5 等等，反之，则可取 1.82、1.85、1.88、1.9 等。这其中所含的“人为”因素非常大，准确度当然不够。因此，在古热流史的研究中，我们应探求一种较科学的计算方法。这些问题在地球热力学和地球化学结合方法中得到了有效的解决。

（三）地球热力学温史计算过程（单井正演）

温度史计算和地史模拟同步进行，设某单井共分为 m 套地层，从下到上的各层底界地质年龄（距今时间，百万年）为 $t_1,t_2,\cdots,t_m,t_{m+1}$表示目前时刻。将地层按正演法从下到上依次编号为 $1\sim m$。该井温度史模拟按以下几步进行：

① 确定 m 个地质时刻（$t_1,t_2,\cdots,t_m$）的古大地热流值、古地表温度和岩石热导率等热学参数，古大地热流和地表温度适用于相应地层的沉积时期。

② 取时刻 t_2，对地层按时间步长 Δt 及空间步长 Δz 划分小段及垂向节点，求各小段厚度、孔隙度，按给定的热学参数由温度史模型计算各节点温度。在地质时期 $t_1\sim t_2$，即第一层从开始沉积到沉积结束，按 Δt 可划分若干时间段，对每个 Δt，均按 Δz 划分垂向节点。最终计算出时刻 t_2 时 1 号地层的平均古温度。

③ 取时刻 t_3，对新沉积的地层按时间步长 Δt 及空间步长 Δz 划分小段及垂向节点、求所有小段厚度、孔隙度，按给定的热学参数由温度史模型计算所有各节点温度。最终计算出时刻 t_3 时 1、2 号地层的平均古温度。

⋮　　⋮　　⋮

依次类推。

④ 取时刻 t_m，按上述步骤最终计算出时刻 t_m 时 1、2、…、$m-1$ 号地层的平均古温度。

⑤ 取时刻 t_{m+1}，按上述步骤计算出该时刻所有地层的平均温度，此时即温度史模拟的今天值。

计算整个地层的古温度时，用分布在该地层中的所有节点上的温度的平均值代替。对全区

所有的人工井均重复①～⑤，由各模拟点的结果可得全区的古地温分布情况。

二、地球热力学和地球化学结合方法(结合法)

结合法中所使用的温度史模型和地球热力学方法相同，不同之处在于结合地球化学资料(如镜煤反射率等)建立了古热流史模型，从而使温度史模拟更加合理。本节主要介绍古热流史模型及结合法的过程。

(一) 温度史模型

选用热传导温度史模型，即式(12-61)：

$$T(z,t) = T_0(t) + Q(t)\int_0^z \frac{1}{K(z,t)}\mathrm{d}z$$

其中各符号的含义在式(12-61)中已说明。

(二) 古热流史模型

所谓求古大地热流史，就是确定各地质时期的古大地热流值。如果我们在目前分层情况下，将时间 t 取为各地层底界的地质年龄，那么求古大地热流史就是求这些时刻的热流值。并认为各地层在开始沉积到结束沉积这段时期内热流值不变，即所求出的古热流值是一个时期的热流值。如果盆地内钻有资料较齐全的标准井，则用该标准井各地层底界相应地质年龄时的古热流值代表全盆地相应时期的的古热流值。现以标准井的某地层为例，建立其底界的古大地热流史模型。

1. 地层底界古热流史数学模型

据 Lerche. I(1984)研究，古热流史数学模型的一般形式可归结为：

$$Q(t) = Q_0(1 + \beta \cdot t) \tag{12-66}$$

式中 $Q(t)$——古大地热流值，HFU=μcal/(cm^2·s)；

Q_0——今天的大地热流值，HFU；

β——$Q(t)$与Q_0的关系因子，Ma^{-1}，随埋藏时间而变化；

t——该层底界在埋藏过程中某时刻的地质年代，Ma；

令：

$$t = t_i - \hat{t}$$

式中 t_i——该层底界今天的地质年龄，距今时间，Ma；

$\hat{t}$——该层底界的埋藏时间，Ma。

因而古热流史模型又可表示为：

$$Q(\hat{t}) = Q_0[1 + \beta \cdot (t_i - \hat{t})] \tag{12-67}$$

在盆地模拟中，一般以埋藏时间 $\hat{t}$ 为线索进行计算，即常用式(12-67)，其中地质年龄 t_i 是已知参数。在式(12-67)中的关键是确定不同时刻 $Q(t)$ 与 Q_0 的关系因子 β 的值。

2. 关于 β 的确定

首先确定 β 的取值范围。我们首先假定该层底界古热流值的范围在 0 和 2 倍的 Q_0 之间，这样的假设是合理的。即：

$$0 < Q(\hat{t}) < 2Q_0$$

由式(12-67)知：

$$-1 < \beta(t_i - \hat{t}) < 1$$

在上述不等式中有变量 $\hat{t}$(埋藏时间)，故难以确定 β 的范围。由于我们认为该层从开始沉积到沉积结束这段时期的古热流相同，因此可以假设埋藏时间 $\hat{t}=0$，即该层底界处于地表，上述不等式成为：

$$-\frac{1}{t_i}<\beta<\frac{1}{t_i}$$

由此确定了 β 的取值区间$(-1/t_i,1/t_i)$。如果地层过厚，上述区间的确定是在对埋藏点加密后形成新的分层资料的基础上进行的。

确定了 β 的取值区间后，采用尝试法确定 β 值。在区间$(-1/t_i,1/t_i)$内等间距取 β 的若干个值，对于每一个 β 值，按照一定的准则判定其优劣，选定其中最优者作为 β 的取值。

下面介绍关于判定 β 值优劣的准则。

3. 用镜煤反射率资料确定 β 值

对于在区间$(-1/t_i,1/t_i)$选取的每一个 β 值，可按古热流模型计算出对应的古热流，进一步由温度史模型可计算出相应的古温度。因此，用镜煤反射率资料确定 β 值的问题可转化为用镜煤反射率资料对相应古地温结果的检验问题。如果检验的符合程度越高，相应的 β 值就越合适。

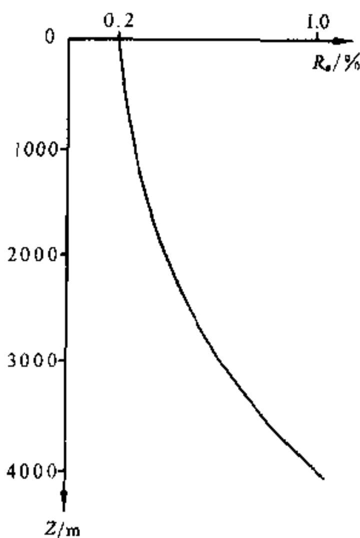


图 12-10 R_o—深度关系曲线

如图 12-10 所示，在 R_o—深度曲线上取 n 个点(R_{ok},z_{ok})，$k=1,2,\cdots,n$。其中地表$(z_{o1}=0)$的镜煤反射率 $R_{o1}=0.2$。结合埋藏史模拟，找出对应于 z_{ok} 的埋藏时间 $\hat{t}_k(k=1,2,\cdots,n)$，其中 $\hat{t}_1=0$，据 Lerche 等人的研究，可以求出关于下面古地温的积分函数的 n 个值：

$$I(\hat{t}_k)=\begin{cases}\int_0^{\hat{t}_k} e^{[T(z,\hat{t})-22]/200}\mathrm{d}\hat{t} & T(z,\hat{t})\geqslant 22\\ 0 & T(z,\hat{t})<22\end{cases}\quad (k=1,2,\cdots,n)\quad(12\text{-}68)$$

理论与实践证明，$I(\hat{t}_k)$与 R_{ok}之间有如下关系：

$$\frac{I(\hat{t}_k)}{\sum_{j=1}^{n}I(\hat{t}_k)}=\frac{(\sqrt{R_{ok}}-\sqrt{R_{o1}})}{\sum_{j=1}^{n}(\sqrt{R_{oj}}-\sqrt{R_{o1}})}\qquad(k=1,2,\cdots,n)\qquad(12\text{-}69)$$

R_o—深度曲线应该满足如下条件：

① 第一个离散点必须是地表$(z_{o1}=0)$，$R_{o1}=0.2$。

② 最后一个离散点的深度应不小于该井最深地层的底界，以满足全井模拟热史的需要。

③ 各离散点的深度值应包括该井各地层的底界，以满足成熟度模型中制作 R_o-TTI 回归曲线的需要。

在计算式(12-68)和式(12-69)时，仅使用 R_o—深度曲线上深度不超过该地层底界的离散点，即下标 k、j 不一定变化到 n。由于古热流值或 β 值选择不当，所模拟出的古温度不完全满足式(12-69)。以均方差 V^2 恒量满足式(12-69)的程度。V^2 越小，满足程度越高。

均方差 V^2 的计算公式如下：

$$V^2=\frac{1}{n(n-1)}\sum_{k=1}^{n}\left(I(\hat{t}_k)\Big/\sum_{j=1}^{n}I(\hat{t}_j)-(\sqrt{R_{ok}}-\sqrt{R_{o1}})\Big/\sum_{j=1}^{n}(\sqrt{R_{oj}}-\sqrt{R_{o1}})\right)^2\quad(12\text{-}70)$$

若在区间$(-1/t_i, 1/t_i)$选取了 l 个 β 值:$\beta_k(k=1,2,\cdots,l)$。对于每个 β_k,按式(12-67)计算出相应的古热流 $Q_k(t_i)$,$k=1,2,\cdots,l$,以 $Q_k(t_i)$为参数进行古温度计算,求出古温度后按式(12-70)求出 l 个均方差,其中与最小均方差相应的古热流值,就是最佳古热流值,而与其对应的 β_k即为最佳选择。此时取 $\beta=\beta_k$,相应的古热流和古温度也同时被认为是最佳的。

由上述方法计算得出的是标准井某地层底界的最佳 β 值,而不同地层底界的最佳 β 值可能不相等,因此可以考虑对不同地层底界取各自最佳的 β 值。另外,由于一个盆地或模拟地区的范围有限,所以对全区同一地层的底界只取一个最佳 β 值。如果全区有多口标准井,每口标准井所计算的最佳 β 值不同,则可取其平均值作为全区该层底界的最佳 β 值。例如:中国东部某坳陷的 6 套地层 Q+N、E_d、E_s^1、E_s^2、E_s^3、E_s^4+K 底界的最佳平均 β 值分别为:

−0.0074、−0.00667、−0.00615、−0.00517、−0.00533、−0.00253

最佳 β 值可以是正值,也可以是负值。当最佳 β 值为正值时,热流从古到今由大变小,古地温梯度也随之由太变小。当最佳 β 值为负值时,热流从古到今由小变大,古地温梯度也随之由小变大。

除了使用镜煤反射率外,我们也可以使用其他地球化学资料。如磷灰石裂变迹径、甾烷、荷烷、芳烃、旋光、花粉等,它们与镜煤反射率起到相同的作用。具体方法如甾烷等比法、荷烷等比法、甾烷的芳烃法、磷灰石面积迹径数密度法、磷灰石迹径长度分布法等(详见《油气盆地数值模拟方法.石广仁.石油工业出版社.1994》)。

(三)结合法模拟古热流史和古地温史步骤(单井)

设盆地内某标准井共分 m 套地层,各层底界的地质年龄分别为:$t_1,t_2,\cdots,t_m$。我们仍以某地层底界(地质年龄$=t_i$)为例说明其热史模拟过程。主要分以下几步进行:

① 求该点的今热流值。按下列公式求取:

$$Q_0 = 10^{-4} K_{t0} \cdot \mathrm{grad}T \tag{12-71}$$

式中 Q_0——该点的今热流值,HFU;

K_{t0}——从地表至该点的沉积物热导率倒数平均值之倒数,$K_{t0}=\left(\frac{1}{z_i}\int_0^{z_i}\frac{1}{K(z)}\mathrm{d}z\right)^{-1}$;

$K(z)$——沉积物热导率;

$\mathrm{grad}T$——实测地温梯度,C/100m;

z_i——该点今天的埋藏深度。

② 求该点的古热流。在区间$(-1/t_i, 1/t_i)$取 l 个 β 值,用式(12-67)求出 l 个古热流值。

③ 求该点的古地温。使用式(12-61)计算出 l 个古地温值。

④ 优选 l 个古地温值和古热流值中的最佳者。利用上面所介绍的方法,在用尝试法求出最佳 β 值的同时,取相应的古热流值和古地温值为最佳选择。

对所有地层重复①~④步。即可求出该标准井各层底界的古热流史和古地温史。

由于在模拟地区大量分布的是人工井点,人工井的分层数据等尚可通过插值法得到,而地球化学资料则不行。对这样的井进行单井模拟时,就利用由标准井求得的古热流史来模拟古地温史。对所有的井点(标准井和人工井)的热史模拟结束后,可得全区古地温的平面分布情况。

热史模拟输出的主要图件包括:单井古热流史图,单井古地温图,剖面古地温图,平面古热流等值图、立体图,平面古地温等值图、立体图等。

三、成熟度史模型

在现今的所有盆地模拟的成熟度模型中,一般采用镜煤反射率的演化史和干酪根产烃率

史来反映生油岩的成熟历史，这主要是因为，在生油量计算过程中，这两项指标可做为重要的参数，而除此以外的所有指标只能表达烃类的成熟度，不能参与生油量的计算，这对盆地模拟意义不大，故盆地模拟中一般只采用镜煤反射率和干酪根产烃率这两项指标。在此，我们只介绍如何计算生油岩的 R_o 史，降解率史的计算将在下一章“生烃史模型”中向大家介绍。要计算 R_o 史，通常要先计算 TTI 史。

（一）TTI 史计算

一般认为，生油母质干酪根的热降解过程符合化学动力学一级反应，即：

$$\frac{dc_a}{dt} = -Kc_a \tag{12-71}$$

式中 c_a——岩石中干酪根含量的浓度；

t——经历的时间；

K——反应速率常数。

反应速率常数 K 可由阿伦尼乌斯方程得到：

$$K = A \cdot e^{-\frac{E}{RT}} \tag{12-72}$$

式中 A——干酪根降解的频率因子，Ma^{-1}；

E——活化能，kcal/mol；

R——气体常数，1.986cal/mol；

T——绝对温度，℃＋273。

无论是 R_o 史还是干酪根产烃率史的计算，都以化学动力学一级反应方程为基础，因此可以说，式(12-71)是成熟度史模型的基础。

假定生油岩在地温间隔 $\Delta T_i = T_i - T_{i-1}$（℃）下经历了 G_i/Ma 地质时间，干酪根浓度从 c_{ai-1} 变化为 c_{ai}，$i=1,2,\cdots n$。当 $\Delta T_i = 10$℃时，反应速度 K_i 可视为常数，因此，对式(12-71)进行逐段积分可得：

$$\begin{aligned}
\ln c_{a1} - \ln c_{a0} &= -K_1G_1 \\
\ln c_{a2} - \ln c_{a1} &= -K_2G_2 \\
\vdots \qquad\quad & \qquad \vdots \\
\ln c_{an} - \ln c_{an-1} &= -K_nG_n
\end{aligned}$$

将上面几个式子相加可得：$\ln c_{an} - \ln c_{a0} = -\sum_{i=1}^{n} K_nG_n$

即

$$\ln c_{a0} - \ln c_{an} = \sum_{i=1}^{n} K_nG_n \tag{12-73}$$

据 Tissot 和 Welte 等人的研究结果，干酪根转化成烃的温度范围界于 50℃～250℃之间，最有利的成油温度为 100℃～110℃，若以 K_t、K_{t+10}、K_{t+20} 分别表示地温为 t℃，$t+10$℃、$t+20$℃时的反应速率，经推导所似有如下关系：

$$1.006 \leqslant \left(\frac{K_{t+10}}{K_t}\right) \Big/ \left(\frac{K_{k+20}}{K_{t+10}}\right) \leqslant 1.086$$

上式表明，在不很严格的情况下，地温每升高 10℃，其反应速率的比值接近一个常数。据韦泊尔斯(Waples. D 1975)研究，该比值可近似取为 2，即：

$$\frac{K_{t+10}}{K_t} \approx \frac{K_{t+20}}{K_{t+10}} \approx 2$$

从该式可以看出，当温度每升高10℃，反应速率K的值增加一倍。用温度间隔100℃～110℃时的反应速率K_{100}去除式(12-73)两端，可得：

$$\frac{\ln c_{a0}-\ln c_{an}}{K_{100}}=\sum_{i=1}^{n} r_i G_i \qquad (12\text{-}74)$$

式中 $r_i=K_i/K_{100}$。一般称式(12-74)右端为温度一时间指数TTI值，即：

$$TTI=\sum_{i=1}^{n} r_i G_i \qquad (12\text{-}75)$$

上式是盆地模拟过程中采用的TTI计算公式。

当温度间隔$\Delta T_i=10$℃时，显然r_i为公比为2的等比级数中的一项，称为温度系数(无因次)。据前人的研究，给出了温度间隔$\Delta T_i=10$℃时r_i的取值见(表12-1)。

表12-1 不同温度下的温度系数

温度范围/℃	温度系数r_i
50～60	0.03125
60～70	0.0625
70～80	0.125
80～90	0.25
90～100	0.5
100～110	1.0
110～120	2.0
120～130	4.0
130～140	8.0
⋮	⋮

在温度史模拟过程中，由于上述公式中的有关参数已能同步得到，故可计算出各生油层在各时期的TTI值演化情况，即TTI史。

在有些盆地模拟系统中，计算TTI史采用以下公式：

$$TTI=\int_0^t 2^{[T(z,t)-105]/10}\mathrm{d}t \qquad (12\text{-}76)$$

式中 t——埋藏时间(Ma)，由地史模型得出埋藏史确定；

$T(z,t)$——古地温，℃，由古地温史确定。

通过上式，可求得各生油层底界的TTI史。式(12-76)和式(12-75)本质上是一致的。

(二) R_o史计算

在计算出了生油层的TTI史之后，进一步要做的工作是找出TTI值与R_o之间的关系，以便根据TTI史来确定R_o史，关于R_o-TTI关系曲线的制作，石广仁提供了一种方法，在TTI值取对数后和R_o是直线关系假设下，将各单井各地层底界TTI的今天值和各层底界的R_o实测值视为观测数据，进行回归得到R_o-TTI关系曲线，R_o的单位是百分数。

一般而言，R_o-TTI关系曲线可以分段表示为：

$$\begin{aligned}
R_o&=a_1\lg TTI+b_1 && 0<TTI\leqslant c_1\\
R_o&=a_2\lg TTI+b_2 && c_1<TTI\leqslant c_2\\
&\vdots && \vdots\\
R_o&=a_{m-1}\lg TTI+b_{m-1} && c_{m-2}<TTI\leqslant c_{m-1}\\
R_o&=a_m\lg TTI+b_m && c_{m-1}<TTI\leqslant c_m
\end{aligned}$$

通过人工观察，确定用m个关系式来描述R_o与TTI之间的关系最为合适后，即可进行分段回归计算，确定相应方程的系数，最终确定R_o-TTI关系曲线。

例如，中国东部某凹陷一口标准井，从R_o一深度曲线知：Q+N、E_d、E_s^1、E_s^2、E_s^3、E_s^4+K六个地层底界的R_o分别为0.24、0.26、0.31、0.43、0.43、0.595，这六个地层底界的TTI今天值分别为：0.12、0.21、0.51、2.58、2.63、10.1，在半对数坐标纸上取TTI为对数坐标，R_o为直角坐标，并将上面的关于R_o和TTI的相应值以坐标形式点在半对数坐标纸上，通过人工观察，发

现该井的 R_o-TTI 曲线只用两个对数表达式表示就可以了，即：

$$R_o = 0.1446\lg TTI + 0.3740 \qquad 0 < TTI \leqslant 2.08$$

$$R_o = 0.3077\lg TTI + 0.3221 \qquad TTI > 2.08$$

用这两个公式可表示该凹陷的 R_o-TTI 关系曲线。中国东部某坳陷和西部某盆地的 R_o-TTI 关系曲线如图 12-11 所示。

除了按上述方法确定 R_o-TTI 关系曲线外，还可使用前人研究的成果，如下所述。

罗泊汀(Lopation. N. V 1971)公式：

$$R_o=1.3011 \cdot \lg TTI-0.5282$$

韦泊尔斯(Waples. D 1976)公式：

$R_o=0.2$　　$0<TTI\leqslant 0.3$

$R_o=(\lg TTI-0.69)/3.82$　　$0.3<TTI\leqslant 10$

$R_o=(\lg TTI-0.14)/2.82$　　$10<TTI\leqslant 30$

$R_o=(\lg TTI-0.67)/1.74$　　$30<TTI\leqslant 75$

$R_o=(\lg TTI-1.01)/1.20$　　$75<TTI\leqslant 300$

$R_o=(\lg TTI-0.59)/0.98$　　$300<TTI\leqslant 2000$

$R_o=(\lg TTI-1.59)/0.73$　　$2000<TTI\leqslant 6000$

$R_o=(\lg TTI-2.09)/0.57$　　$6000<TTI\leqslant 40000$

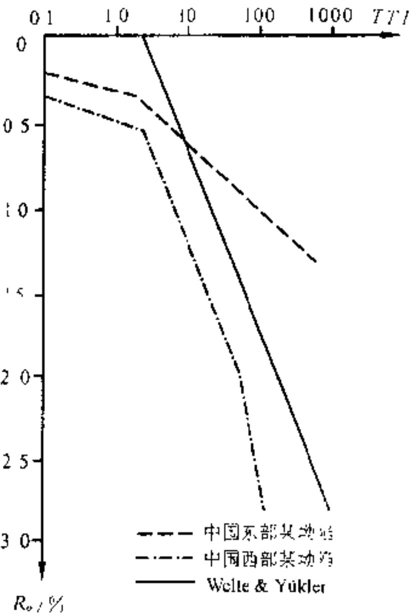


图 12-11　R_o-TTI 关系曲线

上述两个公式在一定条件下是适用的，但实际的模拟结果证明，很多时候计算出的 R_o 史不合理。因此，一般计算 R_o 史，均采用事先制作 R_o-TTI 回归曲线后再计算 R_o 史的方法。该方法由于资料取自模拟区，故所得 R_o-TTI 回归曲线可靠程度相对更高。

在成熟度史模型中，我们是采用了 R_o 和干酪根率产烃做为成熟度指标。TTI 值仅作为计算 R_o 值的过渡参数。为什么不使用 TTI 作为成熟度指标呢？这主要是 TTI 难以反映生油岩的成熟情况，特别是难以确定生油门限。

表 12-2　TTI 和 R_o 生油门限(据矿仁)

Waples 结论与实际情况	生油开始			生油高峰			生油结束		
	深度/m	TTI	R_o/%	深度/m	TTI	R_o/%	深度/m	TTI	R_o/%
Waples. D. W 结论		15			75			160	
中国东部坳陷某区	2700	3	0.5	4000	32	1.0	4600	108	1.3
中国西部凹陷某区	1700	4.1	0.5	3100	77	1.0			
印尼马哈卡姆三角洲		2.4			10			70	

Waples. D. W 的研究认为：生油开始时 $TTI=15$，$R_o=0.5$。生油高峰时 $TTI=75$，$R_o=1.0$。生油结束时 $TTI=160$，$R_o=1.3$。经过对许多地区的实际研究发现，Waples. D. W 的生烃门限绝不是普遍规律。具体见表 12-2。在计算 TTI 值时，所做作的基本假设是"温度每增加 10℃，则烃类成熟反应率增加一倍"，当活化能范围处于 10～20kcal/mol 时，该假设是成立的。但是由于干酪根组分和化学键的复杂性，其活化能的范围远远超出 10～20kcal/mol。TTI 的

第二假设是"把各种类型的有机物视为同一"，当仅考虑煤时，这个假设是成立的。若将它应用于石油的生油岩时，就不一定正确了(Tissot. B. P. 1987)。由此可见，TTI 有其局限性。尽管如此，由于 TTI 仍具有一定的价值及在石油地质研究中多年使用的经验，目前仍在使用 TTI。在 TTI 的使用中必须注意两点：一是不同地区有不同的 TTI 生烃门限。二是不同地区有不同的 R_o—TTI 回归公式。

可见，与 R_o 相比，TTI 值的确很难反映生油门限。因此在盆地模拟中，我们不用 TTI 值做为生油岩成熟度指标，只是用它来求解 R_o 值。

(三) 求 R_o 史的深度回归法

盆地模拟中经常使用一种直接求 R_o 的行之有效的简单方法，从而避免了通过求 TTI 值再求 R_o 的转换计算方法，例如深度回归法。该方法主要步骤如下：

① 结合模拟区资料，求 R_o 与深度 z 的关系曲线(今天)。

② 据地史模型得生油层各时期的埋深。

③ 据关系曲线求生油层的 R_o 史。

该法计算虽较简单，但却较实用，适合连续沉积盆地。

成熟度史(R_o 史)模拟的主要成果图件包括：单井 R_o 史图，剖面 R_o 史图，平面 R_o 等值图、立体图等。

§3 生烃史模型

生烃史模型的主要功能是描述和重建含油气盆地的有机质成熟度史和生烃量史。生烃量计算是油气勘探中的经常性工作，在油气勘探的所有阶段都需要计算生烃量。在盆地模拟系统中，烃类成熟度史和生烃量史的作用不仅在于计算生烃量，而且在于为以后的排烃史和运聚史提供烃类演化环境。

关于成熟度史模型(R_o 史模型)，我们已在"热史模型"一章中介绍过，并且介绍了 R_o 史的计算方法，本章中不再讨论 R_o 史的问题。对于盆地模拟中经常使用的另一成熟度指标——产烃率的演化史，本章中将作为生烃史模型的一部分进行介绍。

计算生烃量史的常用方法有两种：一是 R_o—产烃率曲线法(图版法)，该方法是首先由热模拟实验制作出 R_o—产烃率图版，并根据生油岩的 R_o 史，在图版上确定相应的产烃率史，最终用体积公式法计算出生油岩的生烃量史。该方法一般适用于勘探程度相对较高的地区；二是化学动力学方法，该方法是 Tissot. B. P 根据干酪根的热降解过程符合化学动力学一级反应定律的原理提出的，它把时间、温度因素引入干酪根的热降解过程，在干酪根各键合物质的活化能、对应于每个活化能的各种干酪根所具有的生烃潜量和频率因子等热模拟资料的基础上，模拟干酪根浓度由高到低的变化，求出产烃率史，最终用体积公式法计算出生油岩的生烃量史。该方法一般适用于勘探的各个阶段。上述两种方法都可归结为对产烃率史的求解，因此也可以说，求解产烃率史有上述两种方法。

用上述方法求产烃率史时最好采用模拟地区的实际热模拟资料，在模拟区有关资料无法获取的情况下，可考虑用类比法借用相邻地区或其他类似地区的资料。上述两种方法都是建立在干酪根热降解实验的基础上，属于有机生油理论。

生烃史模型以地史和热史模型的计算结果为基础。使用由前面模型计算出的温度史等结果。从计算结果看，生烃史模型在目前的盆地模拟中是最为重要的。并且生烃史模型的精度直

接影响到排烃史和运聚史模型。

生烃史模型应该包括以下两个子模型：

① 生烃量模型。计算生烃量，传统的体积法模型。

② 产烃率史模型。计算产烃率史(二种方法)，为生烃量模型提供参数。

一、生烃量史模型

在目前的盆地模拟系统中，计算生烃量的基础模型是传统的体积法模型：

$$Q = S \cdot H \cdot C_0 \cdot \beta \cdot \rho_s \tag{12-77}$$

式中 Q——生烃量；

S——生油岩面积(已知)；

H——生油岩中暗色泥岩厚度；

C_o——原始有机碳含量；

β——产烃率，小数；

ρ_s——暗色泥岩密度。

由于在盆地模拟中我们要计算的是生烃量历史，因此我们可将 Q、β 视为地质时间 t 的函数，即上述模型转化为：

$$Q(t) = S \cdot H \cdot C_0 \cdot \beta(t) \cdot \beta_s \tag{12-78}$$

式(12-78)即是生烃量计算中使用的基本数学模型。另外，暗色泥岩密度 ρ_s 随时间的变化不明显，且对计算结果的影响不大，为了简化计算，可将其视为常数(ρ_s=2.33g/cm^3)。否则，应按下列公式计算 ρ_s 随时间的变化：

$$\rho_s(t) = \varphi(t)\rho_w + [1 - \varphi(t)]\rho_r \tag{12-79}$$

式中 t——时间；

$\rho_s(t)$——t 时刻暗色泥岩密度；

$\varphi(t)$——t 时刻暗色泥岩孔隙度；

ρ_w——孔隙流体密度；

ρ_r——暗色泥岩骨架密度。

在式(12-78)中，S、H、C_0 及 ρ_s 均可通过一定途径得到，关键是产烃率 $\beta(t)$ 的求解。求产烃率 $\beta(t)$ 一般有二种方法：即 R_o—产烃率关系曲线法(图版法)和化学动力学方法(Tissot. B. P)。

对单井进行模拟时，如果考虑单位面积的生烃量计算，则式(12-78)变为：

$$Q(t) = H \cdot C_0 \cdot \beta(t) \cdot \rho_s \tag{12-80}$$

由上式计算出的生烃量 $Q(t)$ 称为面积含义下的生烃强度。如果令 $H=1$，则式(12-80)计算出的生烃量称为体积含义下的生烃强度。若未加特别说明，生烃强度是指前者。生烃强度是盆地模拟中的重要结果数据和绘图数据，根据各单井的生烃强度史，平面上可绘制某生油层某时期的生烃强度等值图，反映模拟区某生油层某时期的生烃中心。也可绘制某时期所有生油层的总生烃强度等值图，反映模拟区某时期的生烃中心(图 12-12)。

二、产烃率史模型

(一) R_o—产烃率关系曲线法(图版法)模型

由成热度史模型，可得生油岩的成热度史(R_o 史)。根据由干酪根热模拟实验所得出的 R_o—产烃率关系曲线，针对各生油层的干酪根类型，可求得干酪根的产烃率史。R_o—产烃率关系曲线如图 12-13 所示。

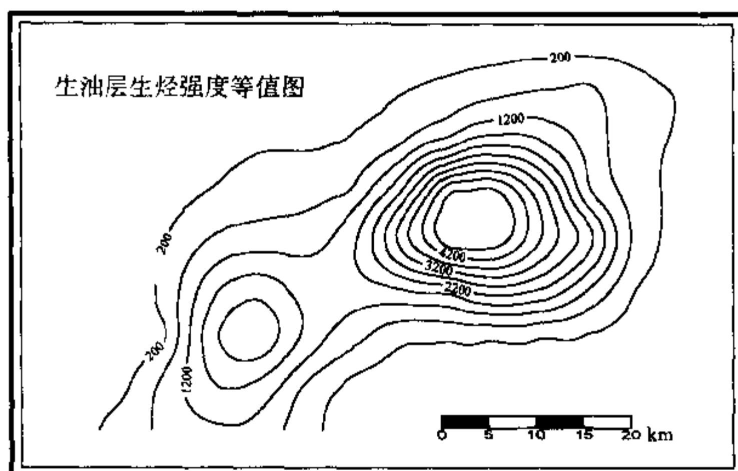


图 12-12 某凹陷生烃强度等值线图

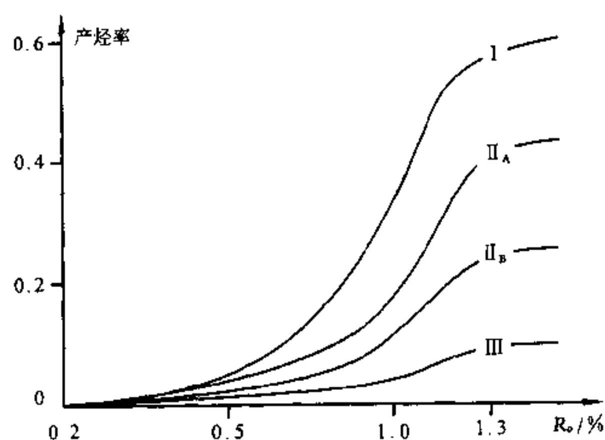


图 12-13 R_o-产烃率关系曲线

如果设地质时刻为 $t_0,t_1,\cdots,t_n,t_n$=目前，地质时刻一般取盆地各地层底界的地质年龄(包括生油层和非生油层)。有机质的成熟度 R_o 取值分别为：

$$R_o(t_k),R_o(t_{k+1}),\cdots,R_o(t_n)$$

其中 t_k 是某生油层结束沉积的地质时刻。从 R_o—产烃率图版中可查取相应的干酪根产烃率值，分别表示为：$\beta(t_k),\beta(t_{k+1}),\cdots,\beta(t_n)$ 。这就是所求出的产烃率史。在盆地模拟系统中，一般是从 R_o—产烃率图版上对各类型干酪根按一定密度采集若干个点作为 R_o—产烃率对应数据，对于不同的 R_o 值采用线性插值方法求相应的产烃率值。

(二) 化学动力学法产烃率史模型

1. Tissot 干酪根热降解数学模型

Tissot 认为，干酪根在温度和时间的作用下向烃类转化的过程可分为二个阶段，即干酪根(A)→降解的中间产物(B)→中间产物(C)。从油气的生成过程考虑，中间产物被认为是液态烃(油)，而最终产物就是天然气。这样，干酪根的热降解生油过程就可划分为成油、成气两大阶

段，如图 12-14 所示。

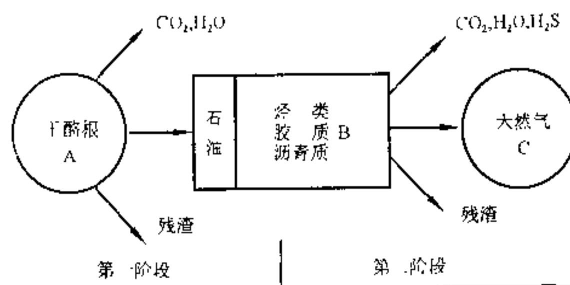


图 12-14　干酪根降解阶段示意图

由 A→B→C 的过程可用这下面的一组符号表示(图 12-15)：

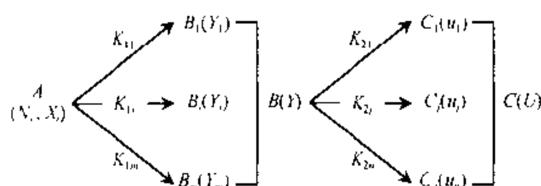


图 12-15　干酪根降解反应流程示意图

式中　N_i——t 时刻干酪根第 i 类键合数目；

X_i——t 时刻干酪根第 i 类键合物质数量；

Y_i——干酪根中第 i 类键合物质裂解产生液态烃数量；

u_j——液态烃(Y)进一步裂解产生 j 型气 C_j 的数量(若认为仅生成甲烷，$j=1$)；

K_{1i}——第 i 类键合物质裂解由 X_i 生成 Y_i 的反应速率；

K_{2j}——液态烃(Y)进一步裂解产生 C_j 的反应速率。

Tissot 认为，干酪根由 6 类不同键合的物质构成，6 类键合的物质降解为石油，进一步降解为气是六个平行的一级反应。因此，可用下面一组方程近似描述干酪根的降解过程：

$$\begin{cases} -\mathrm{d}X_i/\mathrm{d}t = K_{1i}X_i \\ \mathrm{d}u_j/\mathrm{d}t = K_{2j}Y \\ Y = \sum_{i=1}^{6} Y_i \\ \sum_{i=1}^{6} X_{i0} + \sum_{i=1}^{6} Y_{i0} + \sum_{j=1}^{n} u_{j0} = \sum_{i=1}^{6} X_i - \sum_{i=1}^{6} Y_i + \sum_{j=1}^{n} u_j \\ X_0 + Y_0 + U_0 = X + Y + U \end{cases} \quad (i=1,2,\cdots,6 \quad j=1,2,\cdots,n) \tag{12-81}$$

其中反应速率 K_{1i} 和 K_{2j} 可由阿伦尼乌斯方程计算得到：

$$K_{1i} = A_{1i}\exp\left(-\frac{E_{1i}}{RT}\right) \tag{12-82}$$

$$K_{2j} = A_{2j}\exp\left(-\frac{E_{2j}}{RT}\right) \tag{12-83}$$

式(12-81)中第 1 个方程体现了干酪根降解随时间变化的函数关系，第 2 个方程用于求解由干酪根降解产物 Y(液态烃)生成气的数量(生气率)，第 3 个方程表示液态烃总量，第 4 个方

程为物质平衡方程，由该方程可以计算干酪根的产烃率(产油率+产气率)。

方程(12-81)、(12-82)、(12-83)中各种符号的意义如下：

i——第 i 类键合 $(i=1,2,\cdots,6)$；

j——由 Y 生成气体的类别(若认为仅生成 CH_4，取 $j=1$)；

t——时间，Ma；

K_{1i}、K_{2j}——反应速度，Ma^{-1}；

A_{1i}、A_{2j}——频率因子，Ma^{-1}；

E_{1i}、E_{2j}——活化能，kcal/mol；

R——气体常数，1.986cal/mol；

T——绝对温度，℃+273；

Y——生油量(生油率)，克碳/克有机碳；

u_j——生气量(j 型气生气率)，克碳/克有机碳；

X_{i0}——时间为 0 时，干酪根中第 i 类键合物质的初量，克碳/克有机碳；

Y_{i0}——时间为 0 时，干酪根中第 i 类键合物质产生液态烃初量，克碳/克有机碳；

u_{j0}——时间为 0 时，j 型气的初量，克碳/克有机碳；

求解偏微分方程组(12-81)，可选用四阶龙格—库塔方法(Runge—Kutta)。由方程组(12-81)所求出的结果是 t 时刻 $X_i(t)$、$Y(t)$ 以及 $u_j(t)$ 的值。其中 $X_i(t)$ 是干酪根第 i 类键合物质从初始浓度 X_{i0}(初始生烃潜力，小数)经历了 t 时间的热降解后的浓度，$X_i(t)$ 的值小于 X_{i0} 的值，其差值反映了干酪根第 i 类键合物质的产烃率，将 6 类键合物质的产烃率相加，可得干酪根的总产烃率。我们可以利用 Tissot 提供的上述方程组求解生油岩中干酪根的产烃率史(产油率、产气率)，然后在体积法模型的基础上，求生油岩的生烃量史(生油量史、生气量史)。

2. 用 Tissot 方法求烃率史的步骤

如果设生油层埋藏时间按模拟步长确定了 n 个地质时刻：t_0，t_1，$t_2\cdots$，t_k，t_n=目前，考虑到 Tissot 方法将干酪根的热降解划分为生油、生气两个阶段，将求产油率、产气率史的过程归纳为以下步骤：

① 数据准备。确定生油岩中干酪根类型、相应活化能分布及生油潜量数据，干酪根活化能分布及生油潜量数据(见表 12-3)。若模拟地区有这方面的实验资料，则应以实际资料为准。另外，设第二阶段由液态烃生成的气态产物仅为 CH_4，则液态烃生气的反应速率 K_2 由阿伦尼乌斯方程(12-83)计算得到，相应的活化能 E_2 一般取 60～80/kcal/mol，频率因子 A_2 视干酪根类型取与 E_2 相应的值。干酪根热降解阶段依据 R_o 史确定，生油阶段及生气阶段的界限一般取 $R_o=1.3$，R_o 的单位为百分数。

② 取时刻 $t=t_1$，由 R_o 史确定干酪根当前的降解阶段，若 $R_o<1.3$，则当前处于生油阶段，由温史模型确定当前温度，并求各键合物质的降解速率 K_{1i}，用龙格—库塔方法求方程组(12-81)，得该时刻的各类键含物质的残余量 X_i→各类键含物质的生油率→干酪根总生油率，此时，生烃率=生油率。若 $R_o>1.3$，则当前处于生气阶段，由温史模型确定当前温度，并求液态烃降解生气的反应速率 K_2，用龙格—库塔方法求方程组(12-81)，求该时刻干酪根生气率，此时，生烃率=生油率+生气率。

⋮ ⋮ ⋮ ⋮

③ 取时刻 $t=t_k$，重复步骤(2)，求当前干酪根的生油率、生气率和生烃率。

⋮ ⋮ ⋮ ⋮

依次类推。

④ 取时刻 $t=t_n$(今天)，重复上述做法，求今天干酪根的生油率、生气率和生烃率。

⑤ 对所有生油层重复①～④，可求得各生油层中干酪根的产烃率史。

表 12-3　三类干酪根活化能分布及生油潜量(据 Tissot. B. P)

活化能		干酪根类型					
种类	平均值	Ⅰ型		Ⅱ型		Ⅲ型	
E_{1i}	/kcal/mol	X_{i0}	A_{1i}	X_{i0}	A_{1i}	X_{i0}	A_{1i}
E_{11}	10	0.024	4.75×10^{4}	0.022	1.27×10^{5}	0.023	5.20×10^{3}
E_{12}	30	0.064	3.04×10^{16}	0.034	7.47×10^{16}	0.053	4.20×10^{16}
E_{13}	50	0.136	2.28×10^{26}	0.251	1.48×10^{27}	0.072	4.33×10^{25}
E_{14}	60	0.152	3.98×10^{30}	0.152	5.52×10^{29}	0.091	1.97×10^{32}
E_{15}	70	0.347	4.47×10^{31}	0.116	2.04×10^{35}	0.049	1.20×10^{33}
E_{16}	80	0.172	1.10×10^{34}	0.120	3.80×10^{35}	0.027	7.56×10^{31}
$X_0=\sum X_{i0}$		0.895		0.695		0.313	
Y_0		0.051		0.035		0.018	

对于产烃率史模拟，绘制的图件主要包括：各时期生油层产烃率平面等值图、单井产烃率图(结合埋藏史图)、各时期剖面产烃率图(结合剖面图)。

从上面的模拟计算过程中我们可以看出，Tissot 实际上是把干酪根的热降解过程分为生油和生气二大阶段，在生油阶段内没有气生成，在生气阶段中没有油生成。这显然不能全面反映干酪根热降解生烃过程。因此，根据的化学动力学生油原理，对 Tissot 的模型可以做一些适当的拓展。

3. 改进的干酪根热降解数学模型

Tissot 模型是对干酪根生烃机理的简要概括，从模拟角度说比图版法更加合理，但在该模型中并未考虑到油气生成的多阶段性，特别是未考虑到干酪根在生成油的同时也直接生成一定的气，并且生成的气态物除烃外，还有诸如 CO_2、H_2O、N_2 等非烃类气态产物。基于此点，可以进一步改进 Tissot 模型。

在 Tissot 模型的基础上，可以把干酪根的热降解过程分为三个阶段(如图 12-16所示)，对不同的降解阶段用不同的动力学方程组进行描述。第一阶段是生物化学生气阶段(R_o<0.5，R_o 单位是百分数，下同)，该个阶段中的主要产物是 CH_4。没有液态烃产生。第二阶段是热催化生

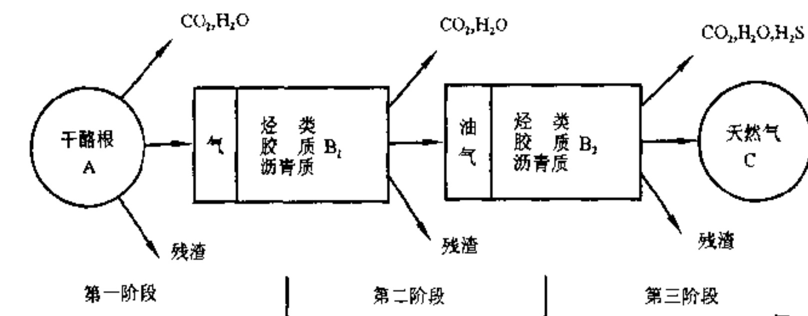


图 12-16　改进的干酪根降解解段示意图

油气阶段(0.5≤R_o≤1.3),该阶段的主要产物是液态烃和CH_4等气体,即干酪根在降解成液态烃的同时也生成一部分气。第三阶段是高温生气阶段(R_o>1.3)。该阶段内液态烃进一步降解生成天然气,而干酪根也直接生成部分天然气,主要产物为CH_4。另外,在各阶段中还伴生CO_2、H_2O、N_2、SO_2、H_2S等气态产物,在上述阶段划分下,可将反映各阶段干酪根热降解过程的化学反应动力学方程组归纳如下:

(1) 生物化学生气阶段(R_o<0.5,主要产物为CH_4等气体)

$$\begin{cases} -\mathrm{d}X_i/\mathrm{d}t = \left(\sum_{j=1}^{6} K'_{2j}\right) X_i \\ X = \sum_{i=1}^{6} X_i \\ Y = Y_0 \\ U = X_0 - X \end{cases} \tag{12-84}$$

该阶段产烃率即是产气率(产烃率=U)。其中K'_{2j}是干酪根直接生气的反应速率。

(2) 热催化生油气阶段(0.5≤R_o≤1.3,主要产物是液态烃和CH_4等气体)

$$\begin{cases} -\mathrm{d}X_i/\mathrm{d}t = \left(K_{1i} + \sum_{j=1}^{6} K'_{2j}\right) X_i \\ \mathrm{d}Y_i/\mathrm{d}t = K_{1i} X_i \\ X = \sum_{i=1}^{6} X_i \quad Y = \sum_{i=1}^{6} Y_i \\ \mathrm{d}U/\mathrm{d}t = \left(\sum_{j=1}^{6} K'_{2j}\right) X \\ U = X_0 + Y_0 + U_0 - X - Y \end{cases} \tag{12-85}$$

(3) 高温生气阶段(R_o>1.3,主要产物是CH_4等气体)

$$\begin{cases} -\mathrm{d}X_i/\mathrm{d}t = \left(\sum_{j=1}^{6} K'_{2j}\right) X_i \\ -\mathrm{d}Y_i/\mathrm{d}t = (K_{2j}) Y_i \\ X = \sum_{i=1}^{6} X_i \quad Y = \sum_{i=1}^{6} Y_i \\ \mathrm{d}U/\mathrm{d}t = \left(\sum_{j=1}^{6} K'_{2j}\right) X + \left(\sum_{j=1}^{6} K_{2j}\right) Y \\ U = X_0 + Y_0 + U_0 - X - Y \end{cases} \tag{12-86}$$

上述方程组中:j——气体的类别(j=1,2,…6);

K'_{2j}——干酪根直接生成j类气的反应速率,Ma^{-1};

K_{2j}——液态烃生成j类气的反应速率,Ma^{-1};

其余各符号的含义和Tissot模型相同。

关于生气的反应速率K_{2j}和K'_{2j},通常的做法是只计算出产生甲烷的反应速率K_{26}和K'_{26},由此推出其他气态产物的反应速率。S. Cao和I. Lerche等人研究认为,不同气态化合物的反应常数与它们的相应分子量有关,较轻气态化合物所需的反应能量较小,反应速率相对较高,反应速率与该气态化合物的分子量成反比。例如:计算出CH_4的反应速率K_{26},可推算产生CO_2的反应速率K_{21}=(16/44)·K_{26},其中16、44分别为CH_4和CO_2的分子量。据此整理出6

种气态产物的反应速率和 CH_4 反应速率之间的关系如下：

$$CO_2: K_{21}=(16/44)\cdot K_{26} \qquad K'_{21}=(16/44)\cdot K'_{26}$$

$$H_2O: K_{22}=(16/18)\cdot K_{26} \qquad K'_{22}=(16/18)\cdot K'_{26}$$

$$N_2: K_{23}=(16/28)\cdot K_{26} \qquad K'_{23}=(16/28)\cdot K'_{26}$$

$$SO_2: K_{24}=(16/64)\cdot K_{26} \qquad K'_{24}=(16/64)\cdot K'_{26}$$

$$H_2S: K_{25}=(16/34)\cdot K_{26} \qquad K'_{25}=(16/34)\cdot K'_{26}$$

计算 K_{26} 和 K'_{26} 时使用阿伦尼乌斯方程，相应的 E_{26} 和 E'_{26} 一般取在 60～80kcal/mol 之间。另外也常取 $E_{26}=E'_{26}$。

4. *动力学方程组的数值积分解法(改进后的模型)*

对动力学方程组(12-84)～(12-86)，可采用龙格一库塔方法求解。更简单的求解过程是采用数值积分方法。归纳方程(12 84)～(12 86)，我们发现其中的常微分方程仅包括两种基本形式：

$$dA/dt=-K(t)A \text{ 或 } dB/dt=K(t)C(t)$$

对这两种类型的常微分方程可以使用数值积分方法求解。将上述两个方程分离变量后分别可得：

$$\frac{dA}{A}=-K(t)\cdot dt \qquad dB=K(t)\cdot C(t)\cdot dt$$

两式两边分别取定积分得：

$$\int_{A_0}^{A}\frac{1}{A}dA=-\int_{t_0}^{t}K(t)dt \qquad \int_{B_0}^{B}dB=-\int_{t_0}^{t}K(t)\cdot C(t)dt$$

即：

$$A=A_0\exp\left(-\int_{t_0}^{t}K(t)dt\right) \qquad B=B_0+\int_{t_0}^{t}K(t)\cdot C(t)dt$$

很显然，对于上面二个方程，我们只要选择一种合适的数值积分方法求出其中定积分的值，那么，求解化学反应动力学方程的问题就解决了。例如，可用辛普森法、梯形公式法或矩形公式法等。

如果设生油层埋藏时间按模拟步长确定了 n 个地质时刻：$t_0,t_1,t_2\cdots t_n,t_n$＝目前，上述定积分的积分上限 t 取其中之一 $(t=t_k)$。则求定积分梯形公式法的计算公式如下：

$$\int_{t_0}^{t}K(t)dt=0.5\sum_{i=1}^{k}[K(t_{i-1})+K(t_i)](t_i-t_{i-1}) \tag{12-87}$$

$$\int_{t_0}^{t}K(t)\cdot C(t)dt=0.5\sum_{i=1}^{k}[K(t_{i-1})C(t_{i-1})+K(t_i)C(t_i)](t_i-t_{i-1}) \tag{12-88}$$

上式中定积分的下限值 $t_0=0$。

5. *用改进的模型求产烃率史的步骤*

仍然

假设生油层埋藏时间按模拟步长确定了 n 个地质时刻：$t_0,t_1,t_2\cdots,t_n,t_n$＝目前，则模拟步骤按以下几步进行：

① 参数准备。除确定和 Tissot 模型相同的参数外，还要确定由干酪根直接生气的反应速率 K'_{2j} 和由液态烃生气的反应速率 $K_{2j}(j=1,2,\cdots,6)$。

② 取时刻 $t=t_1$，根据 R_o 确定干酪根当前的降解阶段，对不同的降解阶段使用不同的动力

学方程组计算生气率、生油率、生烃率 。

⋮　⋮　⋮　⋮

③ 取时刻 $t=t_k$，重复步骤②，求当前干酪根的生油率、生气率和生烃率。

⋮　⋮　⋮　⋮

依次类推。

④ 取时刻 $t=t_n$（今天），重复上述做法，求今天干酪根的生油率、生气率和生烃率。

⑤ 对所有生油层重复①～④，可求得各生油层中干酪根的产烃率史。

最后，需要说明的是所求得的产烃率值是累计量。

三、工区生烃量史计算步骤

在求出模拟地区各人工井中各生油层的产烃率史基础上，可对工区进行生烃量史的计算，计算步骤如下：

① 计算所有平面人工井各生油层的生烃强度（各时期的累计量和净量）。用体积法模型计算各井生油层的累计生烃强度 $Q_{ij}(t)$：

$$Q_{ij}(t) = H \cdot C_0 \cdot \beta(t) \cdot \rho_s \tag{12-89}$$

式中　i——人工井号（$i=1,2,\cdots,m$）；

j——生油层号（$j=1,2,\cdots,l$）；

t——时间（$t=t_0,t_1,t_2\cdots,t_n,t_n=$目前）；

其余符号含意同式(12-78)。

在计算出累计生烃强度基础上再计算阶段净生烃强度，$t_k \sim t_{k+1}$ 的阶段生烃强度：

$$\Delta Q_{ij} = Q_{ij}(t_{k+1}) - Q_{ij}(t_k)$$

② 计算各生油层的生烃量（各时期的累计量和净量）。累计生烃量 $Q_j(t)$：

$$Q_j(t) = \sum_{i=1}^{m} Q_{ij}(t) \cdot \Delta S_i \tag{12-90}$$

式中　j——生油层号（$j=1,2,\cdots,l$）；

m——人工井点数；

ΔS_i——人工井 i 控制的面积；

t——时间（$t=t_0,t_1,t_2\cdots,t_n,t_n=$目前）。

在计算出累计生烃量的基础上再计算阶段净生烃量，阶段 $t_k \sim t_{k+1}$ 的净生烃量：

$$\Delta Q_{ij} = Q_j(t_{k+1}) - Q_j(t_k)$$

③ 计算全区所有生油层的总生烃量（各时期的累计量和净量）。总累计生烃量 $Q(t)$：

$$Q(t) = \sum_{j=1}^{l} Q_j(t) \tag{12-91}$$

式中　l——生油层数；

t——时间（$t=t_0,t_1,t_2\cdots,t_n,t_n=$目前）。

在计算出所有生油层的总累计生烃量的基础上再计算阶段净生烃量，$t_k \sim t_{k+1}$ 的阶段生烃量 ΔQ：

$$\Delta Q = Q(t_{k+1}) - Q(t_k)$$

① 图件绘制。主要包括各种等值线图和直方图。

等值线图包括：

某沉积阶段某一生油层生烃（油、气）强度平面等值图、立体图（阶段量）

某沉积阶段所有生油层总生烃(油、气)强度平面等值图、立体图(阶段量)

某地质时刻某一生油层生烃(油、气)强度平面等值图、立体图(累计量)

某地质时刻所有生油层生烃(油、气)强度平面等值图、立体图(累计量)

直方图包括：

某沉积阶段各生油层生油量直方图(阶段量)

某地质时刻各生油层生油量直方图(累计量)

某一生油层各沉积阶段生油量直方图(阶段量)

某一生油层各地质时刻各生油层生油量直方图(累计量)

所有生油层各沉积阶段生油量直方图(阶段量)

所有生油层各地质时刻各生油层生油量直方图(累计量)

图 12-17 和图 12-18 列出了二种类型的直方图。

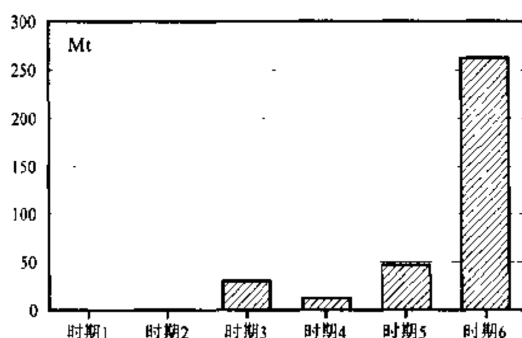


图 12-17 某生油层生烃量史直方图(阶段量)

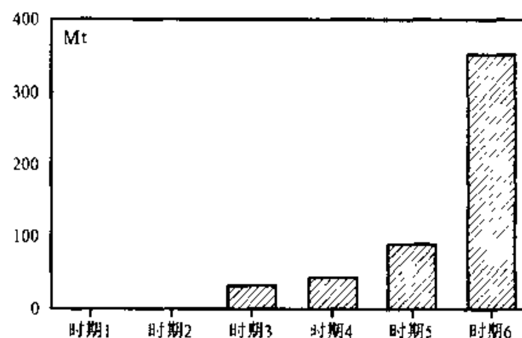


图 12-18 某生油层生烃量史直方图(累计量)

四、改进的体积法生烃量史模型(据石广仁)

(一) 已知 R_o%-产烃率曲线计算生油量和生烃量史

$$E_o(t)=\frac{10^{-12}}{R_{o1}-R_{o2}}\int_{R_{o1}}^{R_{o2}}(z_2-z_1)\cdot M\cdot\rho_s\cdot C_0\cdot\beta(t)\mathrm{d}R_o \tag{12-92}$$

式中 $E_o(t)$——t 时刻该生油层的生烃强度(某人工井单位面积生烃量)；

R_{o1}——t 时刻生油层顶界的 R_o 值，%，由 R_o 史确定；

R_{o2}——t 时刻生油层底界的 R_o，%，由 R_o 史确定；

z_1——t 时刻生油层顶界的深度，m，由地史模型确定；

z_2——t 时刻生油层底界的深度，m，由地史模型确定；

M——生油层内暗色泥岩的百分含量小数；

ρ_s——生油层内暗色的密度，t/km^3；

C_0——原始有机碳含量，小数，据残余有机碳和恢复公式求得；

$\beta(t)$——t 时刻生油岩中干酪根产烃率，小数。

用该公式可以进行生油量和生烃量史的计算，关键是积分上下限的适当的选择。

1. 生油强度及生油量计算

如果我们设 R_o 的生油开始和生油结束的门限分别为 VR_{o1}和 VR_{o2}，相应的深度为 V_{z1}和 V_{z2}，则在用数值方法求式(12-92)中定积分时，有效积分区间为$[VR_{o1},VR_{o2}]$。积分的上下限处

理如下：

① 当 $R_{o1}<R_{o2}<VR_{o1}$时，$E_o(t)=0$，生油岩未开始生油。

② 当 $R_{o1}<VR_{o1}$且 $VR_{o1}\leqslant R_{o2}\leqslant VR_{o2}$时，以 VR_{o1}代替 R_{o1}，V_{z1}代替 z_1，这种情况说明生油层上界尚未进入生油期，有部分生油岩未生油 。

③ 当 $VR_{o1}\leqslant R_{o1}\leqslant VR_{o2}$且 $R_{o2}>VR_{o2}$时，此时说明生油层上界进入生油期，下界超过生油期，故应以 VR_{o2}代替 R_{o2}，以 V_{z2}代替 z_2。

④ 当 $R_{o1}<VR_{o1}$且 $R_{o2}>VR_{o2}$时，说明生油层上下界均未在生油期内，以 VR_{o1}代替 R_{o1}，VR_{o2}代替 R_{o2}，V_{z1}代替 z_1，V_{z2}代替 z_2。

⑤ 当 $R_{o2}>R_{o1}>VR_{o2}$时，$E_o(t)=0$ 。

⑥ 当 $VR_{o1}\leqslant R_{o1}<R_{o2}\leqslant VR_{o2}$时，积分上下限不变。

式(12-92)是对生油层的某种单一类型的干酪根而言的，往往不适合生油层内多种类型干酪根共存的情况。设有 m 类干酪根，各类干酪根的生油强度为 $E_{ok}(t)$ $(k=1,2,\cdots,m)$，则该生油层生油强度可取为：

$$\bar{E}_o(t)=p_1E_{o1}(t)+p_2E_{o2}(t)+\cdots+p_mE_{om}(t) \tag{12-93}$$

式中 $\bar{E}(t)$——t 时刻某人工井生油层生油强度的加权平均值；

p_k——第 k 类干酪根在总干酪根中所占比例，小数，$k=1,2,\cdots,m$，且有：$\sum\limits_{k=1}^{m}p_k=1$。

生油层总生油量(累计)可按下式计算：

$$Q(t)=\sum_{i=1}^{n}\bar{E}_{oi}(t)\cdot\Delta S_i \tag{12-94}$$

式中 $Q(t)$——t 时刻生油全区层的累计生油量，t；

$\bar{E}_{oi}(t)$——t 时刻在井点 i 处生油层的生油强度，t/km²；

ΔS_i——井点 i 所控制的面积，km²；

n——井点数。

2. 生烃强度及生烃量计算

设 R_o 的生烃门限为 VR_{o1}，相应的深度为 V_{z1}，则式(12-92)中定积分的上下限处理如下：

① 当 $R_{o1}<R_{o2}<VR_{o1}$时，$E_o(t)=0$，生油岩未成熟。

② 当 $R_{o1}<VR_{o1}$且 $R_{o2}>VR_{o1}$时，以 VR_{o1}代替 R_{o1}，以 V_{z1}代替 z_1。

③ 当 $VR_{o1}\leqslant R_{o1}<R_{o2}$时，积分上下限不变。

设 m 类干酪根的生烃强度分别为 $E_{ok}(t)$，$k=1,2,\cdots,m$，则该井点生油层生烃强度应取为：$\bar{E}_o(t)=p_1E_{o1}(t)+p_2E_{o2}(t)+\cdots+p_mE_{om}(t)$。而生油层总累计生烃量为：

$$Q(t)=\sum_{i=1}^{n}\bar{E}_{oi}(t)\cdot\Delta S_i \tag{12-95}$$

式中 $Q(t)$——t 时刻生油层全区的累计生烃量，t；

$\bar{E}_{oi}(t)$——t 时刻在井点 i 处生油层的生烃强度，t/km²；

ΔS_i——井点 i 所控制的面积，km²；

n——井点数。

(二) 已知 R_o—产气率曲线计算生气强度和生气量史

在式(12-92)的基础上，考虑到天然气的计量单位，用下式计算生气强度：

$$E_g(t) = \frac{10^{-15}}{R_{o1} - R_{o2}} \int_{R_{o1}}^{R_{o2}} (z_2 - z_1) \cdot M \cdot \rho_s \cdot C_0 \cdot \beta_g(t) \mathrm{d}R_0 \tag{12-96}$$

式中 $E_g(t)$——t 时刻该井点生油层的生气强度，单位面积生气量，$10^8 m^3/km^2$；

$\beta_g(t)$——t 时刻生油岩中干酪根产气率，m^3/t。

其余符号意义同式(12-92)。

用上式进行生油层生气强度计算时，和计算生油强度的情况一样，应注意生气门限的处理。我们设 R_o 的生气开始和结束的门限分别为 VR_{o1}和 VR_{o2}，相应的深度为 V_{z1}和 V_{z2}，积分的有效区间为$[VR_{o1},VR_{o2}]$，则式(12-96)中定积分的上下限处理如下：

① 当 $R_{o1}<R_{o2}<VR_{o1}$时，$E_g(t)=0$，生油岩未生气。

② 当 $R_{o1}<VR_{o1}$且 $VR_{o1}\leqslant R_{o2}\leqslant VR_{o2}$时，以 VR_{o1}代替 R_{o1}，V_{z1}代替 z_1。

③ 当 $VR_{o1}\leqslant R_{o1}\leqslant VR_{o2}$且 $R_{o2}>VR_{o2}$时，以 VR_{o2}代替 R_{o2}，以 V_{z2}代替 z_2。

④ 当 $R_{o1}<VR_{o1}$且 $R_{o2}>VR_{o2}$时，VR_{o1}代替 R_{o1}，VR_{o2}代替 R_{o2}，V_{z1}代 z_1，V_{z2}代 z_2。

⑤ 当 $R_{o2}> R_{o1}>VR_{o2}$时，$E_g(t)=0$ 。

⑥ 当 $VR_{o1}\leqslant R_{o1}<R_{o2}\leqslant VR_{o2}$时，积分上下限不变。

设 m 类干酪根的生气强度分别为：$E_{gk}(t)$，$k=1,2,\cdots,m$，则该井点生油层生气强度应按下式计算：

$$\bar{E}_g(t) = p_1E_{g1}(t) + p_2E_{g2}(t) + \cdots + p_mE_{gm}(t) \tag{12-97}$$

式中 $E_g(t)$——t 时刻某井点生油层生气强度的加权平均值；

p_k——第 k 类干酪根在总干酪根中所占比例，小数，$k=1,2,\cdots,m$，且有：$\sum_{k=1}^{m} p_k = 1$。

生油层总生气量(累计)可按下式计算：

$$Q(t) = \sum_{i=1}^{n} \bar{E}_{gi}(t) \cdot \Delta S_i \tag{12-98}$$

式中 $Q(t)$——t 时刻生油层全区的总累计生气量，$10^8 m^3$；

$\bar{E}_g(t)$——t 时刻在井点 i 处生油层的累计生气强度，$10^8 m^3/km^2$；

ΔS_i——井点 i 所控制的面积，km^2；

n——井点数。

据式(12-98)我们可以计算出生油层的生气量史，并绘制生油层各时期的生气量直方图(累计量、净量)。某时期各生油层的生气量直方图(累计量、净量)，研究主要生气时期和主要生气层。

(三) 化学动力学法求生烃强度及生烃量史

在已知化学动力学法求得产烃率史情况下，求生烃强度史可使用下列公式：

$$E_{hc}(t) = \frac{10^{-12}}{\beta_2 - \beta_1} \int_{\beta_1}^{\beta_2} (z_2 - z_1) \cdot M \cdot \rho_s \cdot C_0 \cdot \beta(t) \mathrm{d}\beta \tag{12-99}$$

式中 $E_{hc}(t)$——t 时刻某井点生油层的生烃强度，t/km^2；

β_1、β_2——t 时刻某井点生油层顶、底界的产烃率，小数；

$\beta(t)$——t 时刻干酪根的产烃率，小数，取各类干酪根产烃率的加权平均值；

其余符号定义同式(12-92)。

生油层总生烃量(累计)可按下式计算：

$$Q(t)=\sum_{i=1}^{n}E_{hci}(t)\cdot\Delta S_i \tag{12-100}$$

式中 $Q(t)$——t 时刻生油层的累计生烃量，t；

$E_{hci}(t)$——t 时刻在井点 i 处生油层的生烃强度，t/km²；

ΔS_i——井点 i 所控制的面积，km²；

n——井点数。

在计算出生油层的生烃强度的基础上，可分别求出生油强度 $E_o(t)$ 和生气强度 $E_g(t)$。

1. 生油量的计算

生油强度计算公式为：

$$E_o(t)=\frac{\bar{U}_o}{\bar{U}_{hc}}E_{hc}(t) \tag{12-101}$$

式中 $E_o(t)$——t 时刻生油层的生油强度，t/km²；

$E_{hc}(t)$——t 时刻生油层的生烃强度，t/km²；

$\bar{U}_{hc}$——生油层中各类干酪根产烃率的加权平均值，小数；

$\bar{U}_o$——生油层中各类干酪根产油率的加权平均值，小数。

生油层生油量：

$$Q(t)=\sum_{i=1}^{n}E_{oi}(t)\cdot\Delta S_i \tag{12-102}$$

式中 $Q(t)$——t 时刻生油层全区的累计生油量，t；

$E_{oi}(t)$——t 时刻在井点 i 处生油层的生油强度，t/km²；

ΔS_i——井点 i 所控制的面积，km²；

n——井点数。

2. 生气量的计算

生气强度计算公式为：

$$E_g(t)=10^{-8}\frac{\bar{U}_g}{\bar{U}_{hc}}E_{hc}(t)\cdot C_{og} \tag{12-103}$$

式中 $E_g(t)$——t 时刻某井点生油层的生气强度，$10^8 m^3/km^2$；

$E_{hc}(t)$——t 时刻生油层的生烃强度，$10^4 t/km^2$；

$\bar{U}_{hc}$——生油层中各类干酪根产烃率的加权平均值，小数；

$\bar{U}_g$——生油层中各类干酪根产气率的加权平均值，小数；

C_{og}——油气转换率，一般可取为 960 m³/t。

上述各参数均为已知，故可求得各生油层的生气强度史。

生油层总生气量：

$$Q_g(t)=\sum_{i=1}^{n}E_{gi}(t)\cdot\Delta S_i \tag{12-104}$$

式中 $Q_g(t)$——t 时刻生油层全区的累计生气量，$10^8 m^3$；

$E_{gi}(t)$——t 时刻在井点 i 处生油层的生气强度，$10^8 m^3/km^2$；

ΔS_i——井点 i 所控制的面积，km^2；

n——井点数。

§4 排烃史模型

排烃史模型的功能是描述和重建含油气盆地的排烃量史和排烃方向史。在盆地模拟中，油气初次运移史的作用不仅仅在于计算排烃量和确定排烃方向，而且能够为运移聚集史模型提供烃源基础。排烃又称油气初次运移，即生油层中生成的油气向运载层或储集层的运移。生油层中的油气，最初呈分散状态，其形成具有工业价值的油气藏首先要经过初次运移。由此可见初次运移在油气藏形成过程中的重要性。排烃史模型建立在地史、热史、生烃史模拟基础之上，是盆地模拟技术的重要组成部分，其精度直接影响到运移聚集史模型的精度。

长期以来，对油气初次运移问题的研究程度不够，存在的问题也较多，特别是对于油气初次运移的机理问题的研究尚不成熟。目前提出的主要排烃机理有以下几种：烃与水呈固有相态运移、水溶液运移、扩散运移、烃溶于气中运移等。不过，多数学者认为：液态烃主要是以自己固有的相态运移，气态烃绝大部分是以溶解于水中的状态运移。另外，油气在初次运移中的主要"运载体"是沉积物中的原生水，初次运移的主要动力主要来自于压实作用、水热增压作用、渗流压力作用、粘土矿物脱水作用、毛细管力作用、甲烷及其他烃类气体的作用等。

基于不同的排烃原理，可设计不同的排烃史数学模型。本章中主要介绍三种研究排烃史的方法：压实法、压差法和渗流力学法，其中前两种方法用于计算液态烃的排出历史，第三种方法用于研究油、气、水的排出历史。另外介绍两种研究排烃方向史的方法：法线方向法和达西定律法。

一、压实法排烃模型

压实法基于连续沉积压实排烃的原理，适用于有规律压实的地区，即孔隙度－深度曲线比较正常的地区。该方法仅适合研究排油史。

压实法的基本思路：首先求出生油层的排出系数史，随后计算求排油强度史，进一步计算排油量史。其中的关键是计算排出系数史。

(一) 排出系数史计算

设某井某生油层在生油开始后某时刻 t_k(压实前)的体积和孔隙度分别 V_0 为和 φ_0，在该时刻后任一地质时刻 t_{k+1}(压实后)的体积和孔隙度分别为 V 和 φ。

根据岩石骨架不可压缩的假设，有：

$$V_0(1-\varphi_0)=V(1-\varphi) \tag{12-105}$$

由此可得：

$$V_0=\frac{1-\varphi}{1-\varphi_0}\cdot V \tag{12-106}$$

或

$$V=\frac{1-\varphi_0}{1-\varphi}\cdot V_0 \tag{12-107}$$

据压实平衡原理，生油层压实前后的体积之差就是所排出的流体体积，即：

$$\Delta V=V_0-V$$

即

$$\Delta V=V_0-\frac{1-\varphi_0}{1-\varphi}\cdot V_0=\frac{\varphi_0-\varphi}{1-\varphi}V_0$$

即

$$\Delta V=\frac{\varphi-\varphi_0}{1-\varphi}V_0 \tag{12-108}$$

我们把生油层在 $t_k \sim t_{k+1}$ 期间的排出系数 C_{ex} 定义为：排出的流体体积 ΔV 与压实前的孔隙体积 V_φ 之比，即：

$$C_{ex} = \frac{\Delta V}{V_\varphi}$$

由于

$$V_\varphi = V_0 \cdot \varphi_0$$

因此有：

$$C_{ex} = \frac{\Delta V}{V_0 \varphi_0}$$

即

$$C_{ex} = \frac{\varphi_0 - \varphi}{(1-\varphi)\varphi_0} \tag{12-109}$$

式中 C_{ex}——生油层的排出系数，$t_k \sim t_{k+1}$ 期间，小数；

φ_0——生油层在生油开始后某时刻（压实前）的孔隙度，小数；

φ——生油层在上述时刻后任意时刻（压实后）的孔隙度，小数。

上式是计算压实前后生油层排出系数的一般公式。结合由地史模型算出的埋藏史，可以确定计算生油层排出系数史的计算公式如下：

$$\begin{cases} C_{ex}(t_1) = 0 & t = t_1 \\ C_{ex}(t_k) = \dfrac{\varphi_{k-1} - \varphi_k}{(1-\varphi_k)\varphi_{k-1}} & t = t_k \end{cases} \tag{12-110}$$

$k = 2, 3, \cdots$，直至今天

式中 $C_{ex}(t_1)$——生油层在埋藏时间 t_1 时的排出系数，t_1 = 生油刚开始，$C_{ex}(t_1)$ 设为 0；

$C_{ex}(t_k)$——生油层在埋藏时间 t_k 时的排出系数，小数，$k=2,3,\cdots$今天；

φ_{k-1}——生油层在埋藏时间 t_{k-1} 时的孔隙度，小数，由地史模型求出；

φ_k——生油层在埋藏时间 t_k 时的孔隙度，小数，由地史模型求出。

由式(12-110)，可算出生油层在各个阶段的排出系数(排出系数史)，也可算出生油层在某一时刻的排出系数，进一步可求出生油层的排油强度(单位面积的排油量)。

（二） 排油强度史的计算

使用下列公式计算生油层累计排油强度史：

$$\begin{cases} E_{ex}(t_1) = 0 & t = t_1 \\ E_{ex}(t_k) = E_{ex}(t_{k-1}) + [(1 - S_{oil})\bar{E}_o(t_k) - E_{ex}(t_{k-1})]C_{ex}(t_k) & t = t_k \end{cases} \tag{12-111}$$

$k = 2, 3, \cdots$，直至今天

式中 $E_{ex}(t_1)$——生油层在埋藏时间 t_1 时的排油强度，t_1 = 生油刚开始，$E_{ex}(t_1)$ 设为 0；

$E_{ex}(t_k)$——生油层在埋藏时间 t_k 时的排油强度，$10^4 t/km^2$，累计量；

$E_{ex}(t_{k-1})$——生油层在埋藏时间 t_{k-1} 时的排油强度，$10^4 t/km^2$，累计量；

$\bar{E}_o(t_k)$——生油层在埋藏时间 t_k 时的生油强度，$10^4 t/km^2$；

S_{oil}——生油层中的束缚油饱和度，小数，取 0.1 左右；

$C_{ex}(t_k)$——生油层在埋藏时间 t_k 时的排出系数，小数。

在式(12-111)中，$(1 - S_{oil})\bar{E}_o(t_k)$ 项为可排出的最大累计生油强度($t = t_k$)，$[(1-S_{oil})\bar{E}_o(t_k) - E_{ex}(t_{k-1})]C_{ex}(t_k)$ 项为阶段 $t_{k-1} \sim t_k$ 内的净排油强度。

因此，$t = t_k$ 时的累计排油强度为：

$$E_{ex}(t_k) = E_{ex}(t_{k-1}) + [(1 - S_{oil})\bar{E}_o(t_k) - E_{ex}(t_{k-1})]C_{ex}(t_k)$$

由上式可计算出某井点生油层的排油强度史。在对模拟区内的所有人工井及所有生油层

都进行排油强度史计算的基础上，可以绘制有关等值线图、立体图，用以研究模拟区单一生油层和所有生油层在某地质时刻及各沉积阶段内的平面主要排油区及排油中心等。

可绘制的主要图件包括：

对某沉积阶段：

某一生油层排出系数、排油强度等值线图、立体图(阶段量)

所有生油层总排出系数、排烃强度等值线图、立体图(阶段量)

对某地质时刻：

某一生油层排出系数、排油强度等值线图、立体图(累计量)

所有生油层总排出系数、排油强度等值线图、立体图(累计量)

(三) 排油量史计算

对于某一生油层，各时期的排油累计量可由下式计算：

$$Q_{ex}(t) = \sum_{i=1}^{n} E_{exi}(t) \cdot \Delta S_i \tag{12-112}$$

式中 $Q_{ex}(t)$——t 时刻生油层的排油量，10^4t；

$E_{exi}(t)$——井点 i 处生油层排油强度，10^4t/km^2；

ΔS_i——井点 i 控制的面积，km^2；

n——人工井点数。

对模拟地区所有的生油层均进行上述计算，可得各生油层的排油量史，在此基础上可绘制有关的直方图，用于研究生油层主要排油时期和主要排油层等。

主要直方图包括：

某沉积阶段各生油层排油量直方图(阶段量)

某地质时刻各生油层排油量直方图(累计量)

某一生油层各沉积阶段排油量直方图(阶段量)

某一生油层各地质时刻排油量直方图(累计量)

所有生油层各沉积阶段总排油量直方图(阶段量)

所有生油层各地质时刻总排油量直方图(累计量)

例如中国东部某盆地 Q+R 沉积阶段内及目前时刻各生油层排油量直方图如图 12-19 和图 12-20 所示。某生油层阶段排油量史和累计排油量史直方图如图 12-21 和图 12-22 所示。

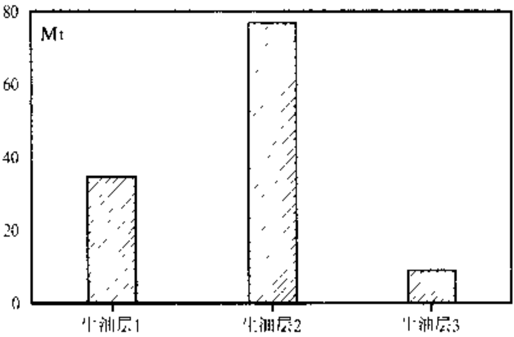


图 12-19 某盆地各生油层排油量直方图
(Q+R 沉积阶段，阶段量)

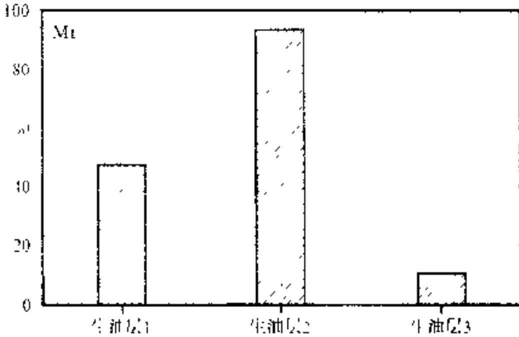
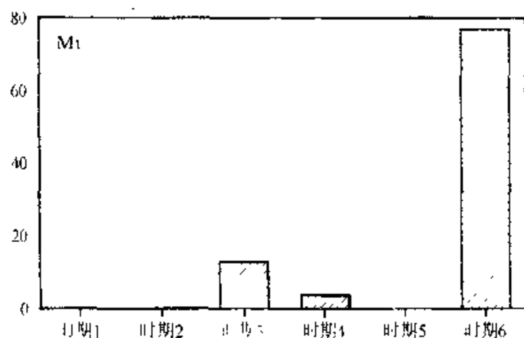


图 12-20 某盆地各生油层排油量直方图
(目前，累计量)



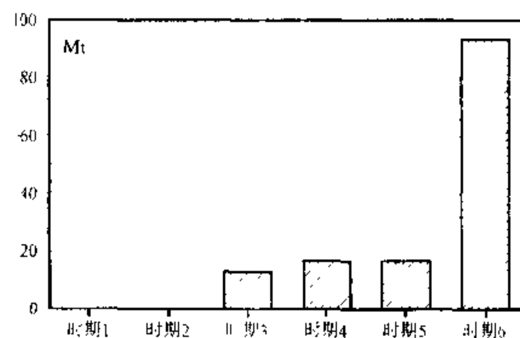


图 12-21　某盆地某生油层排油量直方图（Q+R 沉积阶段、阶段量）

图 12-22　某盆地某生油层排油量直方图（目前，累计量）

上面我们所介绍的压实法仅适用于计算排油量史，不适合于计算排气量史。因为排出系数公式仅适用于液体，不适用于气体。从物理学观点看，油和水的密度随压力和温度的变化不大，而气则会产生很大的变化，故排气不遵循排出系数公式的定义，也就是说，压实法仅适用于计算排油量史，计算排气量史必须另寻它法。

二、压差法排烃史模型

压差法在沉积压实排烃原理基础上，考虑了生油岩与砂岩之间压力差排烃的原理。该方法适用于无规律压实的地区，即孔隙度一深度曲线异常的地区。仅适合研究排油史。

压差法的基本思路和压实法相类似，它也是先求出生油层的排出系数史，然后在此基础上求排油强度史，最终求得排油量史。其中的关键仍然是排出系数史的计算，计算排出系数时，即要仅考虑生油层压实排烃的原理，又要考虑生油岩与砂岩之间孔隙流体压力差排烃的原理，在此基础上的排出系数公式为：

$$C_{ex}(t)=\frac{p_m p_s[\varphi_{cm}\varphi_s(t)-\varphi_{cs}\varphi_m(t)]+p_s\varphi_{cs}[p_m\varphi_m(t)+p_s\varphi_s(t)]S_o(t)}{[p_m\varphi_m(t)+p_s\varphi_s(t)][p_m\varphi_{cm}+p_s\varphi_{cs}S_o(t)]}\qquad(12\text{-}113)$$

式中　$C_{ex}(t)$——单井某生油层在 t 时刻的排出系数，小数；

p_s——生油层的砂岩含量，小数；

p_m——生油层的泥岩含量，小数；

φ_{cs}——生油层进入生油门限时的砂岩孔隙度，小数，根据地史模型和成熟度模型求得；

φ_{cm}——生油层进入生油门限时的泥岩孔隙度，小数，根据地史模型和成熟度模型求得；

$\varphi_s(t)$——生油层进入生油门限后任一时刻 t 的砂岩孔隙度，小数；

$\varphi_m(t)$——生油层进入生油门限后任一时刻 t 的泥岩孔隙度，小数；

$S_o(t)$——生油层进入生油门限后任一时刻 t 的含油饱和度，由生烃和地史模型求出。

在式(12-113)的推导过程中进行了如下假设：

① 生油层进入生油门限之前，全部为孔隙水所充填。

② 生油层进入生油门限之后，干酪根降解生成的油除少量(约 1.42×10^{-4})吸附在干酪根表面外，绝大部分都进入了生油层孔隙。

③ 在沉积埋藏过程中，生油岩孔隙与砂岩孔隙之间存在动态平衡。计算生油门限之后任一时刻 t 的排出系数时，均假设该时刻以前生油岩与砂岩之间相对封闭，没有发生过排油，直到该时刻才解除封闭状态，并且生油岩向砂岩方向排油。即排油在 t 时刻是"一次"完成的，不

是"多次"完成的。该假设不一定符合实际排油过程，但我们关心的是 t 时刻的排油量，其中间过程怎样对计算来说意义不大。

在上述假设下可推导出排出系数计算公式(12-113)，推导过程略。由式(12-113)计算出的排出系数 $C_{ex}(t)$ 是生油层开始压实时刻至 t 时刻的累计量。因此，在计算上述排出系数史的基础上，计算该生油层的累计排油强度公式如下：

$$E_{ex}(t) = \bar{E}_o(t) \cdot C_{ex}(t) \tag{12-114}$$

式中 $E_{ex}(t)$——单井某生油层时刻 t 的累计排油强度，$10^4 t/km^2$；

$\bar{E}_o(t)$——单井某生油层时刻 t 的累计生油强度，$10^4 t/km^2$；

$C_{ex}(t)$——单井某生油层时刻 t 的累计排油系数，小数。

在计算出单井某生油层排油强度史的基础上，进一步可计算生油层的累计排油量史：

$$Q_{ex}(t) = \sum_{i=1}^{n} E_{exi}(t) \cdot \Delta S_i \tag{12-115}$$

式中 $Q_{ex}(t)$——某生油层在时刻 t 的累计排油量，$10^4 t$；

$E_{exi}(t)$——井点 i 处生油层时刻 t 的累计生油强度，$10^4 t/km^2$；

ΔS_i——井点 i 控制的面积，km^2；

n——井点总数。

关于阶段净排油量的计算可取相邻时刻累计排油量之差。

对各生油层均进行排油强度和排油量史计算，在此基础上可以绘制模拟地区关于生油层排油强度史的等值线图、立体图，用于研究主要排油地区、排油中心等。另可绘制有关直方图、用于研究主要排油层系及各生油层主要排油时期等。

在式(12-113)中，我们是假设由泥岩向砂岩排油，生油层中灰岩的含量为0(泥砂剖面)，若灰岩含量不为0，此时生油岩是泥岩和灰岩，且由泥岩和灰岩向砂岩排油，对这种碳酸岩剖面(砂泥灰剖面)不能直接用式(12-113)计算排出系数，必须进行相应的改动，涉及泥岩运算项之处应考虑为泥岩和灰岩的综合运算项，一般将泥岩含量改为泥岩和灰岩含量，泥岩孔隙度改为泥岩孔隙度和灰岩孔隙度的加权平均值。具体改动如下：设生油层中的灰岩含量为 p_l(小数)，$\varphi_l(t)$ 为生油层进入生油门限后任一时刻 t 的灰岩孔隙度(小数)，φ_{cl} 为生油层进入生油门限时的灰岩孔隙度，则排出系数计算公式(12-113)中：

p_m 改为： $p_m + p_l$

$\varphi_m(t)$ 改为： $(p_m\varphi_m(t) + p_l\varphi_l(t))/(p_m + p_l)$

φ_{cm} 改为： $(p_m\varphi_{cm} + p_l\varphi_{cl})/(p_m + p_l)$

和压实法模型情况类似，在由于油和水的密度随温度和压力的变化很小，而气的密度则变化很大。另外，考虑到气是以溶解于油、水中进行初次运移的，不是仅用简单的压差原理所能描述的。因此，压差法模型只能用于研究排油，不能用于研究排气。

三、渗流力学法排烃史模型

前面所介绍的压实法和压差法只能用于求排油量史，不能求排气量史。用渗流力学法不仅能求排油量史，也能求出排气量史及排水量史。

对于平面上的一个人工井点 i，计算油、气、水排出量的公式分别如下：

$$V_o(t) = C\int_{t_m}^{t} \frac{K \cdot K_o \cdot \Delta S_i \cdot P_o}{\mu_o \cdot L} dt \tag{12-116}$$

$$V_g(t) = C\int_{t_m}^{t} \frac{K \cdot K_g \cdot \Delta S_i \cdot P_a}{\mu_g \cdot L} dt \quad (12\text{-}117)$$

$$V_w(t) = C\int_{t_m}^{t} \frac{K \cdot K_w \cdot \Delta S_i \cdot P_a}{\mu_w \cdot L} dt \quad (12\text{-}118)$$

式中 $V_o(t)$、$V_g(t)$、$V_w(t)$ ——井点 i 处生油层在时刻 t 的排油、气、水量 $10^4 m^3$；

t_m——井点 i 处生油层进入生油门限时的埋藏时间，Ma；

t——井点 i 处生油层在进入生油门限之后任一时刻的埋藏时间，My；

K——井点 i 之处生油层绝对渗透率，D；

K_o、K_g、K_w——油、气、水的相对渗透率，小数；

μ_o、μ_g、μ_w——油、气、水的粘度，cp；

P_a——井点 i 处生油层的古超压，由古超压史确定；

ΔS_i——井点 i 控制的面积，km^2；

L——井点 i 处生油层厚度与相邻运载层（或储集层）厚度的平均值，m；

C——单位转换系数，取 $C=35160$。

K_o、K_g、K_w 可近似认为是各自饱和度的函数（据 Nakayama. K 1983）：

$$K_o=[S_o/(1-S_{wir})]^3$$

$$K_g=[S_g/(1-S_{wir})]^3$$

$$K_w=[(S_w-S_{wir})/(1-S_{wir})]^3$$

式中 S_o——节点 i 处生油层的含油饱和度，小数；

S_g——节点 i 处生油层的含气饱和度，小数；

S_w——节点 i 处生油层的含水饱和度，$S_w=1-S_o-S_g$，小数；

S_{wir}——生油层中的束缚水饱和度，小数，可取 0.90～0.99，据 Nakayama. K。

一般，生油层中束缚水饱和度的范围为 0.15～0.35，束缚油饱和度的范围为 0.20～0.30，束缚气饱和度的范围为 0.05～0.10。由于生油层低孔低渗和亲水的缘故，生油层中束缚水饱和度上升，可高达 0.90～0.99，束缚油饱和度下降，可降至 0.10（据 Nakayama. K），束缚气饱和度也下降。

用式(12-116)～式(12-118)，可以分别求得各井点 i 控制的面积 ΔS_i 下生油层的排油、气、水量史，若除以 ΔS_i，可获排油、气、水强度史。在平面上把所有井点的某生油层排油、气、水量相加，即可得全区某生油层的排油、气、水量史。即：

$$Q_o(t) = \sum_{i=1}^{n} V_{oi}(t) \quad (12\text{-}119)$$

$$Q_g(t) = \sum_{i=1}^{n} V_{gi}(t) \quad (12\text{-}120)$$

$$Q_w(t) = \sum_{i=1}^{n} V_{wi}(t) \quad (12\text{-}121)$$

按上述公式计算出的排油、气、水量均以 $10^4 m^3$ 为单位，为了和压实法和压差法模型的单位保持一致，可将其单位转换为 $10^4 t$（万吨）。

在上述计算的基础上可制作关于排油（气、水）强度史的等值图、立体图及有关直方图。

四、排烃方向史

前面的几节中我们介绍了如何计算排油(气、水)量史，在本节中，我们介绍在油气的初次运移过程中，生油层的排油(气、水)的平面方向历史，又称为排油(气、水)流线史。

下面介绍二种常用的方法：法线方向法、达西定律法。

(一) 法线方向法

前面所说的排烃强度是指单位面积的排烃量(面积含义下排烃强度)，它能计算生油层的排烃量，但不能指示平面排烃方向。例如，某井某生油层的生油和排油条件不如周围井的同一生油层，但由于它的厚度比周围井大得多，致使它的排烃强度比周围井大，然而其单位体积的排烃量肯定比周围井小，我们将单位体积的排烃量(体积概念下的排烃强度)定义为"排烃密度"，而排烃方向一般是由排烃密度大者指向排烃密度小者，所以上述生油层的排烃方向是按"排烃密度"从周围井指向该井。若按面积含义下排烃强度确定其平面排烃方向，就会得出相反的结论。

所谓"法线方向法"，是指在生油层某时期排烃密度等值线图上，按线值的大小在每条等值线上绘制若干"法线箭头"，指示该生油层的平面排烃方向。另外，"法线方向法"也可在生油层古超压等值线图上使用，以表示某时期生油层平面排烃趋势。

1. 排烃密度史

根据排烃密度的定义，排烃密度计算公式如下：

$$E_{erv} = \frac{E_{er}}{z_2 - z_1} \tag{12-122}$$

式中 E_{erv}——某井生油层的排烃密度；

E_{er}——某井生油层的排烃强度；

z_1、z_2——生油层的顶、底界深度。

若考虑时间因素，则式(12-122)成为：

$$E_{erv}(t) = \frac{E_{er}(t)}{z_2(t) - z_1(t)} \tag{12-123}$$

据该式可求出生油层排烃密度史：$E_{erv}(t_1)$，$E_{erv}(t_2)$，…，$E_{erv}(t_n)$。

平面上取各时期所有井点同一生油层的排烃密度，在此基础上绘制各时期该生油层的排烃密度等值图，然后确定平面排烃方向。

2. 排烃方向史的确定

按以下几步进行：

① 计算得各人工井各生油层的排烃密度史。

② 绘制各时期各生油层排烃密度等值线图。

③ 在生油层排烃密度等值图中的每条等值线上按一定间距画法线，方向指向数量级较小的相邻等值线。所得到的图件称为"排烃流线等值图"，该图反映了某时期生油层平面排烃方向。生油层各时期的"排烃流线等值图"反映了该生油层平面排烃方向史。

同理，也可在生油层古超压等值图上按上述方法绘制"法线箭头"，形成古超压平面流线图，指示出某时期该生油层的平面排烃趋势。

(二) 达西定律法

基本思路：在生油层古超压史模拟的基础上，应用达西定律确定平面上各人工井点处的烃类初次运移的方向。

达西定律的物理含义为：一种粘度为 1cp 的流体，通过一个截面积为 $1cm^2$ 的孔隙介质，若该孔隙介质两端的压差为 1atm，流出端的流量为 $1cm^3/s$，则孔隙介质的渗透率定义为 1D（达西）。达西定律的数学表达式为：

$$q=-\frac{K\cdot S\cdot \Delta P}{\mu\cdot L} \tag{12-121}$$

式中 q——流体的流量，cm/s；

K——孔隙介质的渗透率，D；

S——孔隙介质的截面积，cm^2；

μ——流体的粘度，cp；

ΔP——孔隙介质两端的压差，atm；

L——孔隙介质的长度，cm。

若式(12-124)两端同除以 S，可得达西定律的另一表达形式：

$$V=-\frac{K}{\mu}\cdot\frac{\Delta P}{L} \tag{12-125}$$

式中 V——流体的流速，$V=q/S$，cm/s。

通过生油层古超压史的模拟，我们得到生油层某时期在平面上各人工井点处的古超压值。根据达西定律，可确定在每个人工井点上流体流动方向

假设平面上局部的井点如图 12-23 所示。通过点 G_0 的 x 方向和 y 方向的流速 V_x 和 V_y 按达西定律可由下式计算：

$$V_x=-\frac{K}{\mu}\cdot\frac{P_{a2}-P_{a1}}{L_x} \tag{12-126}$$

$$V_y=-\frac{K}{\mu}\cdot\frac{P_{a4}-P_{a3}}{L_y} \tag{12-127}$$

式中 V_x——G_0 处流速的 x 方向分量，cm/s；

V_y——G_0 处流速的 y 方向分量，cm/s；

L_x——网格点 G_1 与 G_2 之间的距离，cm；

L_y——网格点 G_3 与 G_4 之间的距离，cm；

$P_{ai}(i=1,2,3,4)$——网格点 G_i 处的古超压，atm；

K——网格点 G_0 处的渗透率；

μ——网格点 G_0 处的流体粘度。

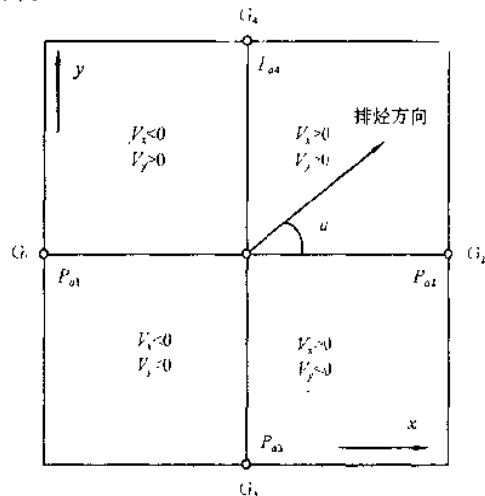


图 12-23 模拟井点古超压分布（某生油层）

在网格点 G_0 处的流体流动方向（排烃方向）用与 x 方向的夹角 α 表示，夹角 α 通过下式确定：

由图 12-23 可知：$\mathrm{tg}\alpha=\frac{|V_y|}{|V_x|}$，故有：$\alpha=\mathrm{arctg}\left(\frac{|V_y|}{|V_x|}\right)$

α——井点 G_0 生油层排烃方向角。

当 V_x、V_y 为正时，用上式可计算出在第一象限的一个角度 α，若 V_x、V_y 其中有一个为负值，则排烃方向角 α 不在第一象限，这时应进行相应的处理，整理如下：

① 当 $V_x,V_y>0$ 时，　$\alpha=\mathrm{arctg}\left(\frac{|V_y|}{|V_x|}\right)$　　　α 在第一象限（0°～90°）。

② 当 $V_x<0, V_y>0$ 时，$\alpha=180°-\mathrm{arctg}\left(\frac{|V_y|}{|V_x|}\right)$　　α 在第二象限(90°～180°)。

③ 当 $V_x<0, V_y<0$ 时，$\alpha=270°-\mathrm{arctg}\left(\frac{|V_y|}{|V_x|}\right)$　　α 在第三象限(180°～270°)。

④ 当 $V_x>0, V_y<0$ 时，$\alpha=360°-\mathrm{arctg}\left(\frac{|V_y|}{|V_x|}\right)$　　α 在第四象限(270°～360°)。

对一个井点做完上述排烃角度 α 的计算之后，可绘制一个表明排烃方向的箭头。对所选定的每个平面上的各人工井点均进行上述计算，即可形成一系列表示流体流动方向箭头，最终得到该时期生油层排液方向流线图(排烃方向)，如图 12-24 所示。对各时期均进行如此处理，可得生油层排液平面方向史。

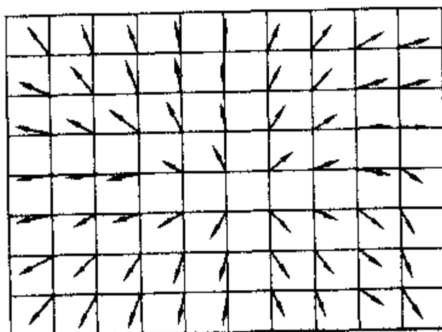


图 12-24　生油层平面排液方向流线示意图

习　题

1. 简述地史模型的目的和构成。
2. 什么是完整的压力史模型？为什么要求地层超压史？正演法中怎样确定超压方程的使用范围？
3. 简述孔隙度史模型的作用及求解过程。
4. 简述古厚度恢复模型的作用及求解。
5. 简述单井地史模拟步骤。
6. 根据单井模拟结果，如何作盆地各地层各时期的等厚图、盆地发育剖面图、埋藏史曲线图、沉积速率等值线图、古压力和异常压力等值线图？
7. 简述回剥技术的思路、关键参数及满足条件和适用范围。
8. 地层骨架厚度公式是什么？怎样选取孔隙度—深度曲线。
9. 地层底界方程是什么？怎样求解？
10. 叙述单井逐层回剥过程(不考虑剥蚀、断层等地质事件)和单井回剥柱状剖面图的绘制。
11. 简述超压技术的目的、主要方程及其作用。
12. 简述古超压方程、古厚度方程及其求解。
13. 简述回剥与超压技术相结合求地史的过程。
14. 为什么要对埋藏点进行加密？怎样加密？

15. 由地史模拟结果可绘制哪些图件？
16. 简述常用的热史模拟方法。
17. 列出常用的温度史模型。
18. 古热流模型的一般形式是什么？。
19. 怎样结合地球化学资料（R_o 资料）确定古热流模型？
20. 简述单井计算古热流、古温度史的步骤。
21. 盆地模拟中所使用的生油岩的成熟度指标有哪几种？
22. 模拟过程中怎样计算地层的镜煤反射率（R_o）史？
23. 由热史模拟结果可绘制哪些图件？
24. 生烃史模型的基本形式是什么？
25. 简述生烃强度的概念和生烃强度的计算方法。
26. 阐述计算生油岩中干酪根产烃率的图版法和求解过程。
27. 阐述计算生油岩中干酪根产烃率的化学动力学法和求解过程。
28. 如何利用图版法计算生油层阶段和累计生烃强度史、生油强度史、生气强度史？
29. 如何利用化学动力学法计算生油层阶段和累计生烃强度史、生油强度史、生气强度史？
30. 由生烃史模拟结果可绘制哪些图件？
31. 叙述排出系数的定义及排出系数史的计算。
32. 叙述生油层排油强度和排油量史计算。
33. 简述压差法排烃模型、含义及适用对象。
34. 如何利用压差法模型计算生油层排油强度史和排油量史？
35. 简述渗流力学法排烃模型、含义及适用对象。
36. 如何利用渗流力学法模型计算生油层排油、气、水量史？
37. 用什么方法确定生油层的平面排烃方向？
38. 由排烃史模拟结果可绘制哪些图件？

第十三章　模拟参数确定及结果分析

模拟参数的确定是盆地数值模拟工作中最重要的环节之一，盆地模拟研究是否成功，很大程度上取决于模拟参数的可信程度。另外，在完成模拟运算后，还必须对有关的模拟结果进行必要的误差检验，以检验模拟结果的计算精度。本章最后还介绍了对模拟结果的分析方法。

§1　主要模拟参数

盆地模拟运算开始之前，做为基础的研究工作，在每个人工井点上必须确定以下几类模拟参数。

一、地质参数

地质参数是模拟的最基本的参数，它的选取必须以实际资料为依据。地质参数主要包括：地层厚度、地层孔隙度、地质年代、剥蚀厚度及岩性组合等。

1. 地层厚度

用钻井分层资料和地震剖面卡取地层厚度，做出全区分层等厚图，如果井较多且平面上分布均匀，也可直接取井的分层资料。也可利用前人的研究成果图件。对于规则的模拟网格，最终必须使用插值的方法将各个地层厚度数据分配到每个人工井点上。

2. 孔隙度

各模拟层的现今孔隙度，可由地震、测井资料经处理后获取，也可利用前人的研究成果，如全区各层孔隙度平面等值线图等。最终分配到每个人工井点上。包括各层的孔隙度一深度曲线（按盆地特征取数条曲线作为代表）。

3. 地质年代与构造运动（剥蚀）

确定各层沉积持续时间、各层的剥蚀厚度及剥蚀开始及结束的时间。

4. 岩性及组合

确定各层的砂、泥岩含量和其它岩性岩石含量，并进行岩性划分。如分为砂岩、泥岩、灰岩。也可进一步细分，如可将各层岩性分为7类：泥岩、碳酸岩、砂岩、砂泥岩、砂质泥岩、含砂泥岩、火成岩。

二、热学参数

1. 大地热流值

确定今大地热流值、古大地热流值。

2. 古地表温度

确定各地质阶段的古地表温度。

3. 岩石热导率

确定不同岩性岩石的热导率（可借鉴标准数据）。

4. 比热

确定岩石骨架和流体比热。

三、有机地球化学参数

地化参数是模拟计算盆地生烃量和排烃量的重要依据，主要包括：生油层暗色泥岩厚度、残余有机碳含量、干酪根类型分布、镜煤反射率及干酪根热模拟实验资料。

1. *有效生油岩厚度*

由钻井取芯和录井资料、统计各层生油岩的厚度。一般定义：残余有机碳含量大于0.5%、颜色深灰或黑色、单层厚度1～30米的纯泥岩，绘制生油岩平面等厚图。最终分配到每个人工井点上。

2. *残余有机碳含量*

包括各生油层残余有机碳含量平面等值图，恢复生油层的原始有机碳含量后，最终分配到每个人工井点上。

3. *干酪根的类型与分布*

包括各油层干酪根平面分布，确定各井点各生油层的干酪根类型。

4. *镜煤反射率*

包括各生油层的镜煤反射率资料，确定 R_o—深度曲线。

5. *热模拟实验*

包括各型干酪根的生烃动力学参数和 R_o—产烃率图版，可由热模拟实验得出或借用（如 Tissot 动力学参数等）。

以上仅列出了盆地模拟中的主要模拟参数，针对不同的模拟地区和模拟要求，还可能进一步增减有关模拟参数。

§2 模拟结果检验

模拟结果检验主要是对正演模拟方法而言的。在单井模拟结束后，得到一系列模拟结果数据，为了检验模拟效果（精度）的好坏，通常要进行以下4个方面的检验。

一、地层厚度检验（对各人工井点）

将模拟得出的各套地层的今天厚度和各地层的实际厚度相比较，若相差过大，则说明地史模拟不成功，需要调整模拟参数之后再重新模拟，直至使两厚度之差在允许的范围之内，通常规定厚度检验相对误差限为±5%。影响模拟地层厚度的主要参数有：地层的原始孔隙度、现今孔隙度、岩性、岩石压缩系数等。

二、孔隙度检验

对各层孔隙度史的今天值与今天的实测值对比检验，通常选取距某标准井最近的人工井（相当于标准井的人工井）的模拟孔隙度史今天值和标准井各地层的实测孔隙度进行对比，相对误差限规定为±5%。影响孔隙度误差的主要参数为：原始孔隙度，岩石压缩系数、岩性等。最后绘制模拟孔隙度和实测孔隙度对比曲线。

三、温度检验

对相当于标准井的人工井进行。将各地层的温度史模拟结果的今天值和标准井实测温度数据进行对比检验，相对误差限±5%。影响温度误差的主要参数包括：岩性、岩石热导率、大地热流值、古地表温度等。最后绘制模拟温度和实测温度对比曲线。

四、成热度检验

对相当于标准井的人工井进行。将各生油层 R_o 史的今天值和标准井实测值进行对比检

验,相对误差限±5% 。影响成熟度检验误差的主要参数包括:R_o—深度曲线、岩石热导率、大地热流值等。最后绘制 R_o 模拟值与实测值对比曲线。

§3 模拟结果综合分析

综合分析是指在盆地模拟结果的基础上,结合其它地质资料,对油气藏起主要控制作用的各种地质因素,如油气源、储层、圈闭、盖层、运移和保存条件等进行系统分析,并对它们之间在时间和空间上的匹配关系进行综合的一套研究方法。该方法主要包括单因素分析、匹配关系分析和综合评价。

一、单因素分析

单因素分析是指对油气藏起主要控制作用的各种地质因素逐一进行独立分析研究。

(一) 油气源分析

油气源分析主要解决:盆地是否有油气生成,油气何时生成,生成多少油气以及在何处生成油气等问题。

1. 油气生成时间分析

分析模拟区各生油层油气开始大量生成的时间。主要依据:生油层生烃量史直方图、所有生油层总生烃史直方图和有关数据表等。

2. 生烃量分析

分析模拟区各时刻和各阶段生油层的生烃量(单一生油层、所有生油层,累计量和阶段量)。主要依据:生油层生烃量史直方图、所有生油层总生烃史直方图和有关数据表等。

3. 主要生烃层系分析

对比模拟地区的所有生油层,分析各地质时刻和目前的主要生烃层,各沉积阶段主要生烃层。主要依据:各地质时刻和各沉积阶段各生油层生烃量(累计量和阶段量)直方图及有关数据表等。

4. 生烃强度及生烃中心分析

在平面上分析盆地各分区生烃强度分布及生烃中心的演化等,确定主要生烃区。并绘制生烃中心评价图。主要依据:各生油层生烃强度等值线图、全区总生烃强度等值线图(立体图)及有关数据表等。

(二) 排烃(初次运移)分析

1. 排烃时间分析

分析模拟区各生油层烃大量排出时间(烃大规模运移时间)。主要依据:各地质时刻和各沉积阶段各生油层排烃量(累计量和阶段量)直方图及有关数据表等。

2. 排烃量分析

确定模拟区各地质时刻和各沉积阶段各生油层的排烃量。主要依据:各地质时刻和各沉积阶段各生油层排烃量(累计量和阶段量)直方图及有关数据表等。

3. 主要排烃层系分析

对比模拟地区的所有生油层,分析各地质时刻和目前的主要排烃层,各沉积阶段主要排烃层。主要依据:各地质时刻和各沉积阶段各生油层排烃量(累计量和阶段量)直方图及有关数据表等。

4. 排烃强度及排烃中心分析

在平面上分析盆地各分区排烃强度分布及排烃中心的演化等，确定主要排烃区。并绘制排烃中心评价图。主要依据：各生油层排烃强度等值线图、全区总排烃强度等值线图（立体图）及有关数据表等。

5. 排烃方向分析

在平面上分析盆地各地质时期各生油层排烃方向，主要依据：生油层排烃流线等值图、生油层排烃平面流线图及有关数据表等。

（三）储层分析

储层分析的涉及面较广，不仅涉及到盆地模拟计算结果，还涉及到沉积相和成岩后生作用的研究。储层分析包括对储层的类型、形态与分布、储集性及成岩后生作用的分析。

1. 储层类型

储层类型按岩性分为碎屑岩（以砂岩为主）碳酸岩、火山岩和变质岩等。

2. 储层形态与分布

研究储层的平面分布及形态演化史，主要依据：盆地模拟中各时期的储层等厚图、沉积相图等。

3. 储集性分析

对储层孔隙度和渗透率进行分析，研究孔隙度和渗透率在空间（平面和剖面上）的变化规律及演化历史。主要对孔隙度进行分析。主要依据：孔隙度一深度曲线、各时期地层孔隙度等值线图及有关数据等。

4. 成岩后生作用

依据地质研究，分析成岩后生作用对孔隙度、渗透率的影响。

（四）圈闭分析

1. 圈闭形成时间

在地史模拟结果的基础上分析构造圈闭的形成时间。主要依据：各地层底界各时期埋深等值线图、沉积发育史剖面图等。另外结合对沉积相带演化规律的研究，确定岩性圈闭的形成时间。

2. 圈闭幅度与面积

分析圈闭的边界并确定圈闭面积与幅度的演化情况。主要依据：各地层底界各时期埋深等值线图（构造史图）、沉积发育史横剖面图等。另外，根据地层孔隙度史等值图、沉积相图等研究砂体的分布范围和厚度史，进一步确定岩性圈闭的面积和厚度的演化情况。

3. 盖层条件及圈闭保存条件分析

分析储层上覆盖层条件的演化情况及圈闭保存条件。主要依据：上覆层泥岩厚度等值图、孔隙度等值图，一般认为当地层孔隙度＜10％时为盖层。另外，各地层底界各时期埋深等值线图不仅适用于研究圈闭形成的时间，也适用于研究圈闭发展（继承性）变化情况，当圈闭发育正常、不被破坏时，保存条件就好，反之，保存条件差。

二、匹配关系分析

主要分析各种地质因素在时间和空间上的相互匹配情况。

（一）时间匹配关系分析

分析生烃时间、大量排烃时间（运移时间）和圈闭形成时间的匹配关系。

1. 有利的匹配关系

圈闭形成时间早于或相同于大量排烃时间，只有这样，圈闭才能捕捉到油气，形成油气藏。

2. 不利的匹配关系

圈闭形成时间晚于大量排烃时间，这时形成的圈闭缺乏足够的油气源供应，难以捕捉到油气并形成油气藏。

（二）空间匹配关系分析

分析生油层、储层、圈闭（包括盖层）的空间位置匹配关系，对圈闭进行评价。

1. 有利的空间匹配关系

纵上生、储、盖连续沉积。没有被不整合面所分隔，生油层与储层两者直接相通。横向上，圈闭距油气源较近。

2. 不利的空间匹配关系

没有良好的生储盖组合，圈闭距油气源较远。

三、综合评价

遵循“构造是主导，沉积是基础，生排是关键，保存是条件”的原则，在单因素分析和匹配关系分析的基础上进行综合评价。

（一）主要生油洼陷评价

根据洼陷各沉积阶段的相对构造运动及沉积特征、洼陷继承性、湖盆发育情况、生油气及排油气能力等各方面条件进行评价，制作主要生油洼陷评价表，将洼陷分为若干类。

（二）构造带评价

根据构造特点、继承性、与油气大量运移时期的匹配、与油气源的匹配及生储盖配置等要素进行评价，制作主要各构造带评价表，将构造带分为若干类。

（三）绘制综合评价图

在综合评价的基础上绘制盆地模拟综合评价图（平面），主要包括：生油中心、生气中心、油气运移方向、最有利区带、次有利区带等。

习 题

1. 盆地数值模拟需要哪几类模拟参数？
2. 简述模拟的地质参数。
3. 简述模拟的地热学参数。
4. 简述模拟的有机地化参数。
5. 简述对正演模拟效果的检验方法。
6. 简述对模拟结果的单因素分析。
7. 简述匹配关系分析。
8. 简述利用模拟结果对盆地进行综合评价的原则及思路。

*第十四章　盆地沉积过程数学模拟简介

§1　前　言

一、概念及发展历史

盆地沉积过程的数学模拟，是指运用沉积学、数学、物理学、流体力学、弹性力学、计算机图形学等学科的知识，在盆地沉积模式的基础上建立合理的数学模型，并以此描述沉积盆地的沉积演化过程，达到从历史角度对盆地形态、地层、沉积相、岩性、物性等方面进行定量研究的目的一门技术。它是为石油的勘探与开发服务的。与含油气盆地数值模拟不同之处在于，盆地沉积过程数学模拟主要描述盆地在不同物源及环境下地层的沉积形成过程及有关特性，而含油气盆地数值模拟则忽略这一点，它主要描述在各地层形成后的有关油气生成、运聚等方面的演化历史。从某种角度上说，盆地沉积过程数学模拟是含油气盆地数值模拟的基础，二者的有机结合将构成完整的盆地数值模拟技术。

盆地沉积过程数学模拟研究起步较早，但发展却较缓慢。这主要是由于地质沉积过程的复杂性和多变性所至。本世纪五十年代，随计算机的出现，一些地质学家开始进行这方面的工作，1960年，美国的哈博(Harbaugh J. W)研究了美国堪萨斯州东南宾夕法尼亚系南辛群灰岩海滩沉积成因问题。1962年，史洛斯(Sloss L. L)结合实际工作著有《勘探中的地层模型》。1966年，哈博建立了复杂的包括海相碳酸盐的沙洲模型，模型的沉积作用是由有机物活动和陆源碎屑屋沉积两部分组成。有机体群落间的互相干扰和它们对海水深度的适应性、碎屑沉积物的影响这二者决定了碳酸盐的沉积作用，另外还考虑了其他一些因素。模型中以相连接的矩形柱表示沉积盆地空间，以不同的柱高表示不同深度的海洋、海岸线、海滩河流三角洲和陆地，模型中还考虑了沉积物搬运、沉积以及构造挠曲作用。1967年，布莱格斯和包莱克(Briggs & Pollack)建立了蒸发岩盆地的数学模型并进行了实际模拟。1970年，奥仑(Allen)建立了陆源碎屑沉积矿床与环境参数(如水深、流速等)之间的关系模型。1970年，哈博发表了具有代表性专著《地质过程计算机模拟》，书中系统论述了模拟流动与搬运的方法，讨论了沉积岩质量模型、沉积盆地模型、蒸发岩盆地模型、三角洲模型、碳酸盐一生态学模型等。1979年，约翰逊(Johson)建立了气一水储集层的模拟模型。

我国在这方面的研究较少。1981年，中国科学院地质所孙惠文等应用哈博的沉积盆地模型对黄骅拗陷进行了初步研究。该模型假设在某时间阶段内，物质通过流水进入盆地，然后按沉积单元逐个沉积。1982年，大庆油田地质勘探开发研究院黄秀桢等人建立了湖泊三角洲沉积数学模型，它假定河床横剖面为矩形，河口是水动力学上的喷嘴，河流入湖形成二维平面射流，所携带的碎屑物是理想颗粒，最终模拟出了不同河湖水动力条件下的三角洲边界、剖面形态和沉积物分选性变化。1989年，江汉石油学院汤军和北京石油勘探开发科学研究院赵旭东在哈博模型的基础上，建立了一套研究沉积盆地沉积过程的数学模拟方法，该方法以Fick扩散定律近似描述物质的搬运(扩散)过程，讨论了沉积盆地中常见的冲积扇和三角洲沉积的数学模型，并对泌阳凹陷双河地区核二段三角洲和核三段冲积扇进行了沉积过程模拟。

综上所述，盆地沉积过程的数学模拟研究到目前虽然取得了一定的进展，但存在的问题也很多，特别是沉积盆地形成过程中沉积机理、沉积条件和沉积环境的复杂性导致了建模的极大困难，致使模拟的实用性不高。今后，随着研究水平和计算机应用水平的不断提高，相信会有进一步的改善。

二、沉积盆地沉积过程模拟的准备工作

对沉积盆地沉积过程进行模拟，首先要建立地质概念模型。地质概念模型反映了地质家对盆地沉积过程进行研究后形成的定性认识，包括对控制盆地沉积的各主要因素的具体认识，这其中包含有盆地沉积过程的普遍规律。地质概念模型形成后，运用系统工程的思想将其系统化，并用数学模型对该地质系统进行近似描述，借助计算机求解数学模型，最终达到定量研究盆地沉积过程的目的。在模拟工作开始之前，应建立沉积盆地沉积过程模拟的动态系统，该系统反映了模拟程序设计的基本思路及各个模块之间的关系。除此以外，还要对模拟地区进行网格划分、模拟参数的确定等工作。对于模拟地区各网格单元中包括水深、物质浓度、沉积厚度等在内的原始状态及其在各个时间阶段的状态还必须以适当的方式进行登记，以便于模拟运算过程中的使用及模拟结果的保存。

1．沉积盆地沉积过程地质动态系统

按系统工程的思路，在地质概念模型的基础上我们将盆地的沉积过程描述为一个动态系统。动态系统把所有的地质作用都理解为该动态系统的不同组分，并它们之间相互联系、相互作用的关系。如物质沉积作用，一方面受水动力、物质浓度、基准面、湖盆形态、坡度、水深、颗粒大小、古气候、物质搬运、湖盆沉降或抬升、构造变动、沉积压实等因素的控制，另一方面沉积作用的变化又影响到水动力条件、盆地形态、水深、物质搬运、沉积压实等，因此可将它们归纳在一个随时间而变化的动态系统内。

对一个沉积盆地来说，我们可建立一个简单的沉积过程动态系统。如图 14-1 所示，它是考

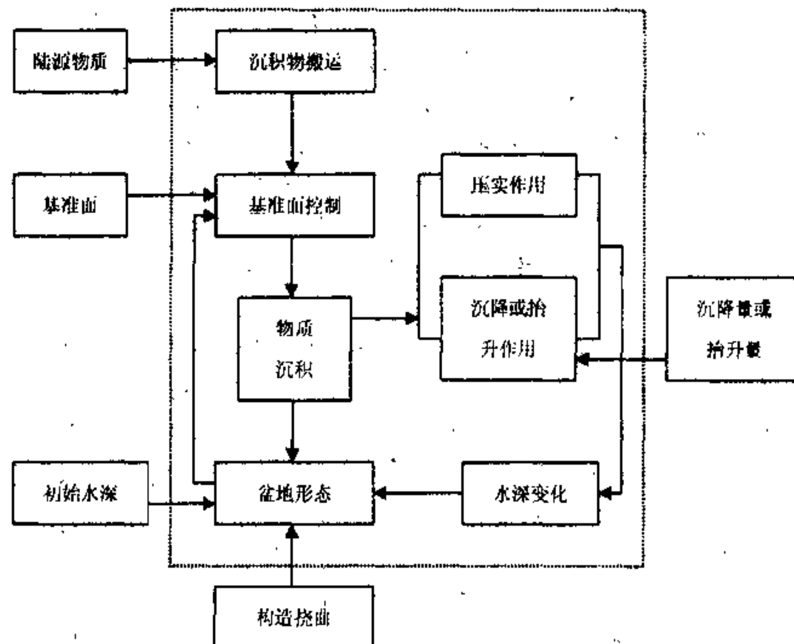


图 14-1　沉积盆地沉积过程动态系统图

虑了陆源物质、沉积物搬运、湖盆形态、湖盆沉降或抬升、构造挠曲、基准面深度等对沉积过程起控制作用的因素在内的动态系统图。从系统外部输入的参数包括基准面深度、陆源物质数量、初始水深、湖盆沉降或抬升量、构造挠曲等，基准面深度控制湖盆沉积物的分布和厚度。陆源物质数量控制着各时间阶段湖盆内参与搬运的物质数量，并通过基准面影响到物质沉积。初始水深决定了盆地的初始形态，通过基准面控制物质沉积。湖盆沉降或抬升、构造挠曲通过改变湖盆形态控制陆源物质的搬运和沉积。在系统内，物质的搬运（扩散）基准面等控制沉积，物质的沉积作用导致湖盆形态的变化，进而又影响到基准面及物质的搬运，随时间的推移最终又影响到物质的沉积，即较早时间的物质沉积影响较晚时间的物质沉积。由此可见系统内各部分之间的相互联系。

2. 模拟网格划分

对于二维沉积模拟和三维沉积模拟，分别需要进行二维网格和三维网格划分。二维模拟是以湖盆的平面或剖面范围为模拟对象，描述湖盆平面或剖面上不同位置的沉积演化史。而三维模拟则是以湖盆存在的空间为对象，进而描述湖盆空间各位置的沉积演化史。考虑到三维模拟的复杂性及计算耗时巨大，而模拟精度并未因此而有明显的提高，因此，沉积过程的模拟一般采用二维模拟方式。二维模拟包括平面二维和剖面二维，与此相应的是平面网格和剖面矩形柱的划分，平面网格划分是指在一个包含盆地平面沉积范围的矩形内对工区进行网格划分。形成大小相同的多个有序相邻网格单元，模拟结果就是以分布于各网格单元内各时期各种物质的沉积厚度表示的。为模拟计算方便，一般使用单位面积的正方形网格，每个网格单元赋予行列编号，如图 14-2 所示，用 i 表示网格单元的行号，$i=1,2,\cdots,n$，从上到下行号由小变大，用 j 表示网格单元的列号，$j=1,2,\cdots,m$，从左到右列号由小变大。

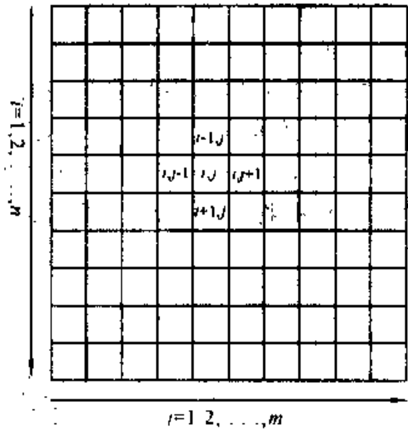


图 14-2 网格类型

网格单元的密度及其实际控制面积据具体情况而定，即要考虑模拟的精度，又要考虑模拟所占用的机时。另外，按网格单元的位置将其分为不同的类型，包括界外（湖盆边界以外）、左上角、上边界、右上角、右边界、右下角、下边界、左下角、左边界、界内（湖盆边界内）共 10 种类型，分别用数字 0～9 表示（如表 14-1 所示）。如有特殊需要也可适当增加类型。通过网格单元类型的设定，应该能看出工区的平面形态（如图 14-3）。

表 14-1 网格单元类型与位置

网格单元类型编号	网格位置
0	湖盆边界外
1	左上角
2	上边界
3	右上角
4	右边界
5	右下角
6	下边界
7	左下角
8	左边界
9	湖盆边界内

0	0	0	0	1	2	3	0	0	0
1	2	2	2	9	9	9	2	2	3
8	9	9	9	9	9	9	9	9	4
8	9	9	9	9	9	9	9	9	4
8	9	9	9	9	9	9	9	9	4
8	9	9	9	9	9	9	9	9	4
8	9	9	9	9	9	9	9	9	4
8	9	9	9	9	9	9	9	9	4
7	6	6	6	9	9	9	9	9	5
0	0	0	0	7	6	6	6	5	0

图 14-3 网格划分及编号

对于剖面二维模拟而言，由于它是在穿过湖盆的一条横剖面上进行模拟，因此不存在平面网格的划分问题，取而代之的是在剖面上按等间距划分若干矩形柱(如图 14-4)，并将这些矩形柱从物源向湖盆中心依次编号，柱子的高度表示水深，它没有类型之分，最终模拟结果是这些矩形柱内的物质沉积厚度史。

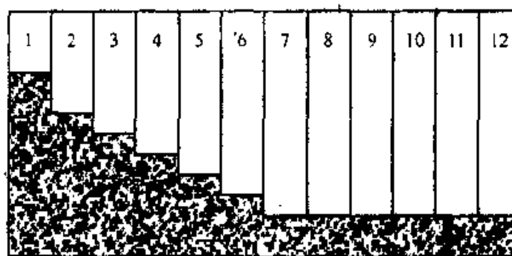


图 14-4 剖面模拟的矩形柱

3. 模拟数据登记法

对平面二维模拟，按上述网格划分方法，所有平面网格单元的组合描述了完整的湖盆平面形态，如果已知各网格单元的沉积演化史，整个湖盆的沉积演化史便可由此得知。模拟运算的实际对象是各个平面网格单元，而模拟数据登记就是将湖盆的初始以及各沉积时期的状态(数据)分配到各个平面网格单元上的方法。模拟过程中必须登记各网格单元的水深、搬运到达的不同类型物质的浓度、不同类型物质的沉积厚度、湖盆沉降量等数据在各沉积时期的变化情况。使用计算机高级语言能方便地以随机文件的形式存储上述各项数据，存储时将各项数据按时间、网格单元、物质粒级的不同分别存储于随机文件的不同“记录”中，这里的“记录”是指计算机磁盘上的数据存储单元。设网格行号 $i=1,2,\cdots,n$，列号 $j=1,2,\cdots,m$，物质粒级总数为 f，l 表示第 l 种粒级，时间阶段 t 时各网格单元的水深、物质沉积厚度的登记情况如图 14-5 和图 14-6 所示，其它数据的登记情况与此类似。

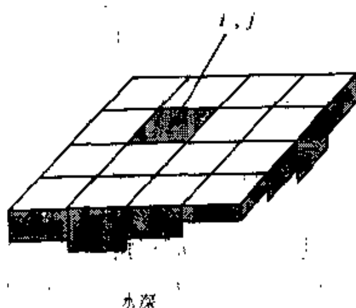


图 14-5 水深的登记(时间 t)

记录号:$(t-1)\times n\times m+(i-1)\times m+j$

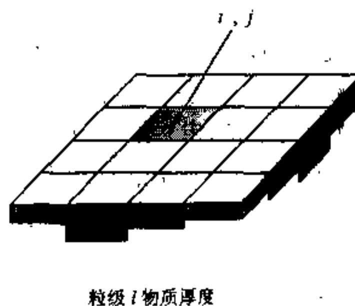


图 14-6 物质厚度的登记(时间 t，粒级 l)

记录号:$(t-1)\times n\times m\times f+(i-1)\times m\times f+(j-1)\times f+l$

对剖面二维模拟，登记方法相对简单一些，它只需要考虑时间，矩形柱号，物质粒级来定义随机文件的记录号，如编号为 j 的矩形柱水深的记录号可设定为 $(t-1)\times m+j$，它表示 t 时间第 j 个柱子的水深，而 $(t-1)\times m\times f+(j-1)\times f+t$ 表示 t 时间第 j 个柱子第 l 种粒级物质的沉积厚度记录号。

§2 模拟模型

盆地内地层的成因是多方面的，即有陆源物质入湖沉积，又有湖水蒸发造成盐类的沉淀、生物体沉积、宇宙尘埃等。在湖盆发育期，各种沉积可能兼而有之。对于不同成因的地层，由于

沉积机理及物源的不同，应采用不同的数学模型进行描述。J. W. Harbaugh 等曾讨论了沉积盆地模型、蒸发岩沉积模型、碳酸岩一生态学模型等。本章在 Harbaugh 盆地模型的基础上，介绍对河流或洪水携带碎屑物质（硅质）进入湖盆后进一步搬运并沉积，形成盆地地层这一演化过程的模拟方法。即对沉积盆地沉积过程的模拟。

沉积盆地沉积过程模拟需要描述的关键问题有二个，一是物质的搬运过程，二是物质的沉积过程。对物质搬运过程的描述必须了解物源的供给情况、物源种类、搬运条件，确定搬运的距离及方向（范围），最终解决各时间阶段碎屑物质到达各个网格单元的数量（浓度），它是沉积的物质基础。在描述物质的沉积时，必须了解控制沉积的主要因素、参与沉积的物质及沉积顺序，最终解决各时间阶段碎屑物质在各个网格单元的沉积量。只有将这些问题搞清楚了，才有可能对盆地的沉积过程进行模拟。

沉积盆地的沉积过程中，河流或洪水携带的陆源碎屑物由某一入口进入湖盆，经搬运后沉积下来。主要的沉积类型包括：冲积扇沉积、三角洲沉积、浊流沉积、滨湖沉积、浅湖沉积、半深湖沉积、深湖沉积等。在模拟过程中，我们将各种类型的沉积描述为不同物源在不同搬运距离以及不同沉积环境下产生的结果。如三角洲沉积，认为是河流携带的泥沙不断进入湖盆，由于坡度减缓，水流扩散，流速降低而在河口区沉积下来。对冲积扇沉积，认为是洪水或河流携带的近源物质在湖盆陡坡带沉积形成，其物源供给是间歇性的。滨湖沉积、浅湖沉积、半深湖沉积、深湖沉积等认为是各种粒级的陆源物质入湖后受其本身的动能、湖盆形态、基准面及湖盆水动力条件等的影响进一步搬运（扩散）并沉积的结果。因此，我们就可以在统一的搬运和沉积概念模型下建立相应的数学模型来描述沉积盆地的沉积过程。对于不同的沉积类型，通过设定盆地各网格单元不同的水深、基准面深度、不同类型物质的扩散速度及浓度、沉积比例等参数来加以区分。以下讨论平面二维模拟的物质搬运模型和沉积模型。

一、物质搬运模型

扩散过程可以看作是空间内物质浓度趋于平衡的过程，例如，我们将一颗高锰酸钾晶体投入有水的烧杯，经溶解与扩散过程，最后整个水中的高锰酸钾浓度将是均匀的。这一过程是在没有流体运动的条件下发生的，扩散是分子随机运动的结果。在模拟过程时，可以有效地运用扩散方法，即使对一个实际过程来说并非严格意义下的扩散过程。例如，水中悬浮物质的搬运可以作为一个扩散过程来对待（据 J. W. Harbaugh）。如果将一些染上颜色的砂粒置于海滩白浪带，在波浪湍流的作用下它们将很快弥散开，如果扩散是造成搬运的唯一原因，结果是产生圆形的弥散，但实际上在水流对流运动的作用下产生了一个非圆形的弥散，这是一个水流作用下的扩散过程。在描述陆源物质入湖后的搬运过程时，我们可以把该过程是看成是物质在水流等因素作用下由入口的高浓度区向湖盆中心的低浓度区扩散的过程。

1. 数学模型

陆源碎屑物进入湖盆后，受本身动能、湖盆水动力、重力等因素的影响继续向湖盆内搬运，将这一过程理解为物质浓度扩散过程的前提下，可使用 Fick 第一扩散定律描述物质的质量通量和浓度随距离的变化率之间的关系：

$$J_x = -k\frac{\partial c}{\partial x} \tag{14-1}$$

式中 J_x——物质的质量通量（单位时间内物质流过单位面积的质量）；

k——扩散系数；

c——物质的浓度；

x——平行于流动方向且垂直于通过面的坐标轴。

类似地可定义 y 轴方向和 z 轴方向的物质质量通量 J_y 和 J_z 分别为：

$$J_y = -k\frac{\partial c}{\partial y} \qquad J_z = -k\frac{\partial c}{\partial z}$$

Fick 第一扩散定律在数学表达方式上类似于描述多孔介质中流体流动的达西定律，但它所描述的扩散过程和流体流动的规律是不相同的。

如图 14-7 所示，假设物质通过一个立方体单元，单元的边长分别为 $\Delta x,\Delta y,\Delta z$，将立方体单元的六个侧面标记为：

l,r——x 方向上的左面和右面；

f,a——y 方向上的前面和后面；

t,b——z 方向上的顶面和底面。

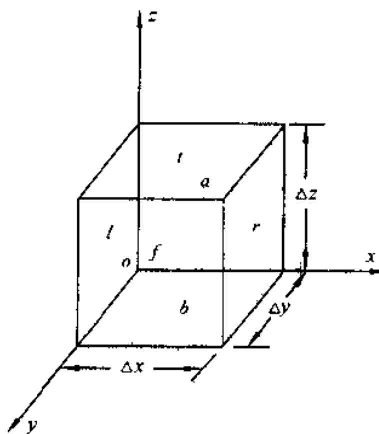


图 14-7　物质通过的立方体单元

如果物质进入并流出这个立方体单元，则基本的质量连续关系为：

$$输入 - 输出 = 积累 \tag{14-2}$$

单位时间内通过立方体左面和右面的物质质量流量分别为：

$$M_l = \Delta y \cdot \Delta z \cdot J_{xl} \qquad M_r = \Delta y \cdot \Delta z \cdot J_{xr}$$

其中：J_{xl} 和 J_{xr} 表示通过立方体左面和右面的物质质量通量，物质质量流量从左面到右面通过距离 Δx 的净差为：

$$\begin{aligned}\Delta M_x &= M_l - M_r \\ &= \Delta y \cdot \Delta z \cdot J_{xl} - \Delta y \cdot \Delta z \cdot J_{xr} \\ &= \Delta y \cdot \Delta z \cdot (J_{xl} - J_{xr}) \\ &= \Delta y \cdot \Delta z \cdot \Delta J_x\end{aligned}$$

即

$$\Delta M_x = \Delta y \cdot \Delta z \cdot \Delta J_x \tag{14-3}$$

类似地可定义 y 方向和 z 方向的净差分别为：

$$\Delta M_y = \Delta x \cdot \Delta z \cdot \Delta J_y \tag{14-4}$$

$$\Delta M_z = \Delta x \cdot \Delta y \cdot \Delta J_z \tag{14-5}$$

对立方体单元，质量流量在时间间隔 Δt 内的净积累可定义为三个方向上净差的代数和再乘以 Δt：

$$积累 = (\Delta M_x + \Delta M_y + \Delta M_z) \cdot \Delta t \tag{14-6}$$

另一方面，质量流量在时间间隔 Δt 内的净积累又可定义为物质在 Δt 起始时的浓度 c_1 和结束时的浓度 c_2 之差再乘以立方体单元的体积：

$$积累 = (c_1 - c_2) \cdot \Delta x \cdot \Delta y \cdot \Delta z = \Delta c \cdot \Delta x \cdot \Delta y \cdot \Delta z \tag{14-7}$$

由式(14-6)和式(14-7)得：

$$(\Delta M_x - \Delta M_y + \Delta M_z) \cdot \Delta t = (c_1 - c_2) \cdot \Delta x \cdot \Delta y \cdot \Delta z \tag{14-8}$$

将式(14-3)～(14-5)代入上式得：

$$(\Delta y \cdot \Delta z \cdot \Delta J_x + \Delta x \cdot \Delta z \cdot \Delta J_y - \cdot \Delta x \cdot \Delta y \cdot J_z)\Delta t = (c_1 - c_2) \cdot \Delta x \cdot \Delta y \cdot \Delta z$$

用 $\Delta x \cdot \Delta y \cdot \Delta z$ 除以上式两端得：

$$\frac{\Delta J_x}{\Delta x}+\frac{\Delta J_y}{\Delta y}+\frac{\Delta J_z}{\Delta z}+\frac{\Delta c}{\Delta t} \tag{14-9}$$

如果立方体单元无限缩小，则上式变为：

$$\frac{\partial J_x}{\partial x}+\frac{\partial J_y}{\partial y}+\frac{\partial J_z}{\partial z}=\frac{\partial c}{\partial t} \tag{4-10}$$

将式(14-1)及类似公式代入得到描述物质的确定性扩散过程的费克第二定律：

$$\frac{\partial c}{\partial t}=-k\left(\frac{\partial^2 c}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}+\frac{\partial^2 c}{\partial z^2}\right) \tag{4-11}$$

式中 c——物质浓度；

k——扩散系数；

x，y，z——沿扩散方向的坐标轴。

如果扩散流的速度随方向不同而变化，则应使扩散系数在三个方向上分别取不同的值，如果我们去掉式(14-11)中正负号并重新定义三个方向的扩散系数为：k_x，k_y，k_z，则方程(14-11)变成：

$$\frac{\partial c}{\partial t}=k_x\frac{\partial^2 c}{\partial x^2}+k_y\frac{\partial^2 c}{\partial y^2}+k_z\frac{\partial^2 c}{\partial z^2} \tag{4-12}$$

上式即是描述物质浓度扩散的三维确定性方程。如果忽略 z 方向上的扩散，可得描述物质浓度扩散的二维确定性方程：

$$\frac{\partial c}{\partial t}=k_x\frac{\partial^2 c}{\partial x^2}+k_y\frac{\partial^2 c}{\partial y^2} \tag{4-13}$$

上式中 k_x，k_y 为 x，y 方向的扩散系数。用上式可近似描述进入湖盆内的碎屑物质随时间的搬运过程。另外，上式适用于平面二维模拟，对于剖面二维模拟，可以采用更简单的方法描述搬运过程。

2. 数值解法

为了便于模拟运算，必须把扩散方程转换成有限差分形式。下面以二维扩散方程为例，讨论其有限差分表达式及边界条件。

二维扩散方程如式(14-13)所示。这个方程的有限差分表达式要求把它看作是展布在 x 轴与 y 轴方向上矩形网格的一个函数。按第二节所介绍的网格划分方法，分别用 i，j 表示网格单元的行和列，并将时间分为一系列足够小的等间距增量 Δt，那么对于网格单元(i，j)有：

$$\frac{\partial c}{\partial t}=\frac{c_{t+1,i,j}-c_{t,i,j}}{\Delta t} \tag{14-14}$$

$$\frac{\partial^2 c}{\partial x^2}=\frac{c_{t,i,j+1}+c_{t,i,j-1}-2c_{t,i,j}}{(\Delta x)^2} \tag{14-15}$$

$$\frac{\partial^2 c}{\partial y^2}=\frac{c_{t,i+1,j}+c_{t,i-1,j}-2c_{t,i,j}}{(\Delta y)^2} \tag{14-16}$$

综合方程(14-13)～(14-16)得：

$$\frac{c_{t+1,i,j}-c_{t,i,j}}{\Delta t}=k_x\frac{c_{t,i,j+1}+c_{t,i,j-1}-2c_{t,i,j}}{(\Delta x)^2}+k_y\frac{c_{t,i+1,j}+c_{t,i-1,j}-2c_{t,i,j}}{(\Delta y)^2} \tag{14-17}$$

如果考虑正方形网格，则 $\Delta x=\Delta y$，方程简化为：

$$c_{t+1,i,j} = c_{t,i,j} + \frac{\Delta t}{(\Delta x)^2}[k_x \cdot (c_{t,i,j+1} + c_{t,i,j-1} - 2c_{t,i,j}) + k_y \cdot (c_{t,i+1,j} + c_{t,i-1,j} - 2c_{t,i,j})]$$

(14-18)

对于已知的网格划分及初始浓度分布，由上式可迭代求出各时间阶段分布在各个网格单元上的浓度值。式(14-18)的解取决于网格单元的边长 Δx、时间间隔 Δt 及各个网格单元的初始浓度值。在求解的过程中还必须规定 $\Delta t/(\Delta x)^2 \cdot k_x$和 $\Delta t/(\Delta y)^2 \cdot k_y$ 的值在 0.0～0.5 之间，否则将导致高浓度骤然降为低浓度，甚至出现负值，这与扩散过程的物理定律是相矛盾的。

二维扩散方程的有限差分表达式(14-18)所包含的算术运算具有平滑浓度值的作用，这与扩散过程本身是相符合的，它所描述的扩散趋势是补偿各网格单元的浓度值，这涉及到 x、y 方向上前后相邻的四个网格单元浓度值的求和运算，对于各角网格单元和边界网格单元，符合要求的相邻网格分别只有二个和三个，对这些网格单元不能按式(14-18)求解，必须在边界条件的限制下改变式(14-18)的形式。

常使用两种边界限制条件，一种是假设在该时间阶段内穿越边界的物质浓度的梯度保持不变，另一种是假设在该时间阶段内穿越边界的物质浓度值保持不变，在此采用第一种边界条件，即：

$$\frac{\partial c}{\partial x} = 0 \qquad \frac{\partial c}{\partial y} = 0$$

以右上角和右边界网格单元为例说明边界的处理情况。如图 14-8(a)所示的右上角网格单元(i, j)，虚设网格单元($i-1$, j)和(i, $j+1$)，对应于上述边界条件有：

$$c_{t,i,j} = c_{t,i,j+1}$$
$$c_{t,i,j} = c_{t,i-1,j}$$

代入式(14-18)并整理得：

$$c_{t+1,i,j} = c_{t,i,j} + \frac{\Delta t}{(\Delta x)^2}[k_x \cdot (c_{t,i,j-1} - c_{t,i,j}) + k_y \cdot (c_{t,i-1,j} - c_{t,i,j})] \qquad (14\text{-}19)$$

对于如图 14-8(b)所示的右边界网格单元(i, j)，虚设网格单元(i, $j+1$)，对应于边界条件有：$c_{t,i,j} = c_{t,i,j+1}$，代入式(14-18)并整理得：

$$c_{t+1,i,j} = c_{t,i,j} + \frac{\Delta t}{(\Delta x)^2}[k_x \cdot (c_{t,i,j-1} - c_{t,i,j}) + k_y \cdot (c_{t,i+1,j} + c_{t,i-1,j} - 2c_{t,i,j})] \qquad (14\text{-}20)$$

另外三个边角的情况可依此类推。图 4-18 右上角和右边界网格单元

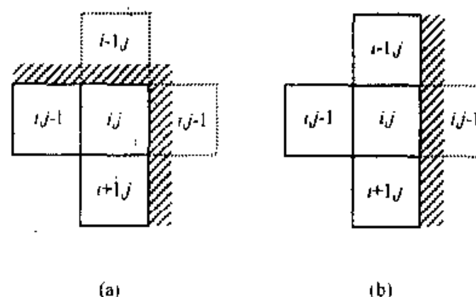


图 14-8　右上角和左边界网格单元

二、物质沉积模型

物质的沉积作用是物质搬运作用的继续，各种粒级的物质经搬运(扩散)后在湖盆某网格单元内具有一定的浓度，受基准面及其它因素的控制，有部分物质沉积下来，形成一定的沉积厚度(可能为 0)，其余部分继续进行搬运(扩散)。物质沉积模型最终求解结果是各时间阶段各网格内各种粒级物质以厚度方式表示的沉积量。下面讨论在物质浓度作用下物质沉积量的计算以及基准面、湖盆沉降、水进水退等因素对沉积的控制作用。

1. 物质浓度下控制的沉积量

在将碎屑物质进入湖盆后继续搬运的过程描述为物质浓度扩散过程的前提下，用物质搬

运模型来描述物质浓度在湖盆内的分布情况。物质浓度反映了沉积物供给量的多少，物质的沉积量和浓度密切相关。当湖盆内某具体位置（某网格单元）的物质浓度已知时，按先沉积粗粒，后沉积细粒的顺序计算出不同粒级物质的沉积量，而且各种粒级物质中只有一部分沉积下来（与沉积系数有关），剩余的物质继续进行扩散。在单位面积（网格单元）内，如果将某粒级物质的浓度理解为到达该网格单元的“重量”，并假设它为 c_v，另设该粒级物质的沉积系数为 c_s，密度为 ρ_s，沉积厚度为 h，则有：

$$c_s \cdot c_v = \rho_s \cdot h$$

因此该粒级物质的沉积厚度 h 可表示为：

$$h = \frac{c_s}{\rho_s} c_v \tag{14-21}$$

由上式可计算出各粒级物质在不同浓度和沉积系数下的最大沉积厚度，这里的“最大”厚度是指不受其它因素控制的沉积厚度。如果有其它因素（如基准面）的控制作用，则实际达不到最大厚度。式(14-21)中物质密度 ρ_s 是已知的，物质的浓度 c_v 是由搬运模型求出的，沉积系数 c_s 是来源于实际的可调试经验值。

2. *基准面（波基面）的控制作用*

由于波浪、水流、物质颗粒大小、盆底坡度和粗糙度等因素的相互影响而产生一个平衡面，即基准面，沉积作用受基准面的控制。各网格单元内在基准面以上不能接受沉积。如图 14-9 所示，设湖盆某网格单元的水深为 d，基准面深度为 b，按式(14-21)计算的最大沉积物总厚度为 L，实际的沉积厚度为 H，在基准面的控制下总有 $H \leqslant L$，剩余部分继续参与搬运。考虑按以下三种情况分别处理基准面对沉积的控制作用：

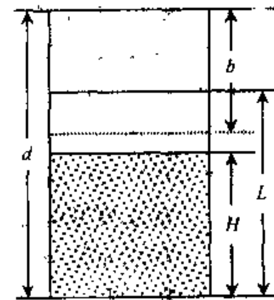


图 14-9 基准面与沉积厚度
d—水深；b—基准面深度；
L—最大沉积厚度；
H—实际沉积厚度

①当网格单元某阶段水底面高于或等于基准面时，任何物质不能沉积，继续向其它位置扩散（搬运）。此时有 $d \leqslant b$，该阶段实际沉积厚度 $H=0$，剩余量为 L。

②当网格单元某阶段水底面界面至基准面的深度不能容纳最大沉积厚度 L 时，则沉积最多至基准面，剩余部分继续向其它位置搬运。此时有 $L \geqslant d-b$，仅有部分沉积，实际沉积厚度 $H=d-b$，剩余量为 $L-(d-b)$。

③当网格单元水底面至基准面的深度能够容纳最大沉积厚度时，则接受全部沉积。此时有 $L \leqslant d-b$，实际沉积厚度 $H=L$；剩余量为 0。

3. *湖盆沉降的控制作用*

盆地发育期湖盆整体沉降造成水深随时间而增加，基准面相对上移，进一步又可接受新的沉积。在模型中通过增加各网格单元水深值的方法描述湖盆的沉降作用。可以采用两种方法改变水深值：第一种方法是给定各网格单元内的水深随时间的一个增量值（常量），这种方法适合于描述湖盆随时间的面产生的整体均匀沉降。第二种方法是用一个简单的比例常数把沉降和沉积作用联系起来，各网格单元的沉降量等于沉积厚度与比例常数之积，这种方法适合于描述水底沉积由于压实作用而产生的形变。特殊情况下，若比例系数为 1，则沉积量等于沉降量，若比例系数为 0，则不出现沉降。用与描述湖盆沉降类似的方法描述湖盆的上升作用，此时湖盆的水深随时间而变浅。

另外，如果使沉积作用落后于沉降作用一定的时间增量。则产生滞后沉积效果。

4. 水进和水退的控制作用

水进和水退作用的结果造成湖盆沉积空间的增大或减少，此时湖盆水深相应发生变化。如果假设湖盆整体呈均匀沉降或上升，也可以通过增加或减少表示湖盆不同位置网格单元的水深描述水进和水退作用。描述水进作用时，控制湖盆沉降量大于沉积量。考虑到水进作用造成湖盆面积的扩大以及水深的增加，湖盆边缘方向水深为0的网格单元随湖盆的沉降作用而具有非0值，而湖盆内网格单元的水深亦随之增加；描述水退作用时，控制湖盆沉降量小于沉积量。从湖盆边缘开始各网格单元的水深随沉积或上升作用渐变为0。

§3 沉积类型与时间序列

如前所述，对于不同的沉积类型，在模拟过程中使用相同的搬运和沉积模型进行描述，它们之间的区别在于网格单元的位置、水深、物质类型、扩散系数以及沉积系数等各项模拟参数。如果假设物源入口区仅有三角洲和冲积扇沉积二种类型，而其它类型沉积是这二种沉积类型的物源向湖盆其他位置继续扩散并沉积的结果，那么可以在模拟过程中只考虑这二种沉积类型。在表示物源入口区的网格单元上，通过不同的水深、物质类型、扩散系数和沉积系数等模拟参数来体现这二种沉积类型的差别。

另外，在模拟运算过程中应该考虑二者之间物源供给方式的区别，这种区别导致沉积时间序列的不同。如对于三角洲（建设性）沉积，河流携带物源以长期不间断方式涌入湖盆，在每个沉积时期内，搬运、沉积过程在时间上是连续的，即物质一边搬运、一边沉积，同时又接受新的物源。模拟的时间序列可考虑为连续、相等、间隔较小的多个时间阶段，这种时间阶段数当然越多越好，一般按模拟精度的要求人为确定其个数。时间阶段数越多，时间增量就越小，模拟的精度就越高，但所耗费的机时也越多。图14-10描述了三角洲模拟过程中时间、物源、搬运沉积之间的关系。对于冲积扇沉积，物源由间歇性河流或洪水携带，集中于若干个沉积时期内进入湖盆，每一沉积时期内进入湖盆的物源经搬运过程并完全沉积下后转入下一沉积时期，并再次接受新的物源。沉积过程在总体上是不连续的，但在同一沉积时期内具有连续性。因此，它的模拟时间序列可考虑分为若干沉积时期，每个沉积时期内再分为若干连续、相等、间隔更小的多个时间阶段，这些时间阶段内没有大量新的物源进入湖盆。沉积时期个数原则上等于洪水期个数，但考虑到洪水的频繁性将导致计算机无法承受的计算量，可将实际的洪水期次数归纳为若干个主要洪水期，即沉积时期。图14-11描述了冲积扇模拟过程中时间、物源、搬运沉积之间的关系。

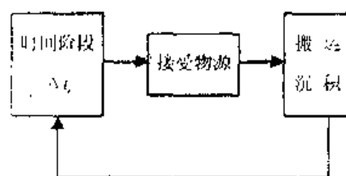


图14-10 三角洲模拟时间序列

三角洲和冲积扇沉积模拟都是盆地沉积过程模拟的组成部分，在考虑其模拟流程有区别的同时，又必须考虑到它们在模拟时间阶段上的统一性，这样才可能描述整个盆地的沉积过程。为此我们将沉积体所经历的总沉积时间均分为许多小的时间阶段，用 M 表示其中若干个时间阶段的

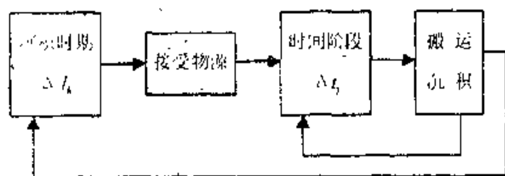


图14-11 冲积扇模拟时间序列

和，并假设湖盆每相隔 M 便接受新的物源。这样我们可以根据 M 值的大小控制这二种沉积类型的模拟时间流程。如图 14-12 所示，当 $M=1$ 时，表示三角洲沉积时间流程，当 $M\gg1$ 时，表示冲积扇沉积时间流程。

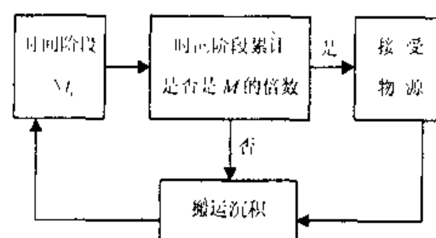


图 14-12　沉冲盆地模拟时间序列

在上述模拟时间序列的基础上，对三角洲和冲积扇沉积分别进行模拟，并将同一时间阶段各网格单元的物质浓度、各类型物质的沉积厚度等累加，求出湖盆在各个时间阶段的沉积厚度及岩性(粒级不同的碎屑)纵向分布。如果盆地内的沉积类型不止两种，则应分别考虑相应的模拟模型以及时间序列，在模拟时间序列统一的基础上，将各网格单元上各沉积类型的模拟结果叠加获得沉积盆地的沉积历史。

§4　剖面上物质的搬运与沉积

哈博(Harbaugh J. W，1970)曾论述了一个简单的沉积盆地模型，这个模型在沿物源入口方向的盆地剖面上描述沉积物的搬运、沉积过程。模拟时将剖面按等间距分为许多柱子(见图 14-4)，每个柱子内的水深值是不同的，另外还假设了各柱子内的沉积系数、沉降量等参数。在任意时间阶段内的搬运沉积过程如下：物源从剖面一侧进入湖盆，然后沿物源方向逐个柱子进行搬运，在每个柱子内都可能发生沉积，搬运到达某柱子内的物质除去沉积下来的数量外，其余部分被搬运到下一个柱子。若每个时间阶段内均有物源进入湖盆，并重复搬运、沉积，其结果是沉积物向湖盆中心依次推进。

表 14-2　各柱内的物质负载及沉积量

柱子序号	物质负载量	沉积量	剩余量
1	S	kS	$S(1-k)$
2	$S(1-k)$	$kS(1-k)$	$S(1-k)^2$
3	$S(1-k)^2$	$kS(1-k)^2$	$S(1-k)^3$
4	$S(1-k)^3$	$kS(1-k)^3$	$S(1-k)^4$
5	$S(1-k)^4$	$kS(1-k)^4$	$S(1-k)^5$
…	…	…	…
n	$S(1-k)^{n-1}$	$kS(1-k)^{n-1}$	$kS(1-k)^n$

物质搬运沉积的规律描述如下：如果水深大于基准面的深度，则到达某一具体柱子的物质部分沉积下来。在水足够深以至于基准面不起作用时，每一种粒级的物质按比例沉积。设进入湖盆的物质总负载量为 S，各柱子内的沉积系数为 k，从物源处开始的第一个柱子的沉积量为 kS，剩余量为 $S-kS$ 或 $S(1-k)$，并全部转移到下一个柱子。第二个柱子的沉积量为 $kS(1-k)$、剩余量为 $S(1-k)-kS(1-k)=S(1-k)^2$，并全部转移到下一个柱子。依次类推，按此规律计算出其余各柱子的沉积量，如表 14-2 所示。上述规律描述了水足够深的情况，如果引入另外的控制因素，则情况会有所不同。例如基准面的控制作用、地壳的沉降作用、水进水退作用等，这时可按类似于平面二维模拟的方法进行处理。

§5 主要模拟参数及成果输出

一、主要模拟参数

模拟过程中需要确定湖盆的初始水深、物源入口及初始浓度、扩散系数、基准面深度、沉积系数、沉降系数、沉积经历的时期、孔—深曲线函数等有关模拟参数。

1. *湖盆的初始水深*

指分配到湖盆各网格单元的原始水深。根据对湖盆的有关研究确定，它反映了湖盆接受沉积之前的原始形态。

2. *物源入口及初始浓度*

在湖盆周边确定某些网格单元为物源入口点，在这些网格单元上给定各种粒级物质的初始浓度(重量)，其余网格单元初始浓度一般赋为0。它反映了各粒级物质开始进入湖盆时的平面初始浓度分布。原则上可考虑任意多粒级的物质，本章的程序中考虑了5种粒级的物质。该项参数需要在模拟过程中不断调整，以使沉积厚度的模拟结果和实际情况基本吻合。

3. *扩散系数*

指分配到湖盆各个网格单元中各粒级物质的扩散系数，它反映了物质在该网格单元内的搬运能力。需要给定各网格单元在x,y方向上的扩散系数来分别描述物质在两个方向上的搬运能力，其值范围在0.0～1.0之间。该参数在各网格单元的不同取值情况反映了物质扩散的均匀与否，当各网格单元内x方向上的扩散系数均取相同值时，反映了物质在x方向上的均匀扩散。当各网格单元内y方向上的扩散系数均取相同值时，反映了物质在y方向上的均匀扩散。而各网格单元内扩散系数取不同值时，反映各粒级物质在平面上的非均匀扩散。扩散系数在模拟过程中也需要不断调整，以使模拟结果和沉积体的平面形态及范围相吻合。

4. *基准面深度*

指分配到各个网格单元上的各种粒级物质的基准面深度，粗粒物质基准面一般较浅，细粒物质基准面相对校深。模拟过程中可能对某些网格单元的基准面深度进行细微的调试，但对模拟结果不产生全局性的根本影响。

5. *沉积系数*

指分配到湖盆各个网格单元中各粒级物质的沉积系数，它反映了到达该网格单元物质总量的沉积百分比。用一定的物理意义解释该参数的含义是困难的，可把它看作是一个经验的、可调试的参数来使用。

6. *湖盆的沉降系数*

指分配到湖盆各网格单元上的随时间的沉降量，反映湖盆底部随每一个时间阶段的沉降量。它的取值不同将导致湖盆形态演化史的不同。

7. *沉积时期数*

指将沉积体经历的总沉积时间划分成的沉积时期个数。每个沉积时期表示一定的时间间隔。确定沉积时期个数时要考虑不同沉积类型的影响。

8. 各粒级物质的原始孔隙度及孔—深曲线

求各时间阶段的沉积物厚度的演化及孔隙度的演化时使用。

以上各项模拟参数存储于多个自定义的计算机磁盘文件的不同记录中。

二、模拟流程及成果输出

模拟主要经过模型建立、网格划分、参数确定、模拟运算、结果对比检验、参数调整、成果输出等步骤。模拟主要流程如图 14-13 所示。

主要输出成果包括工区各时期水深图表、各粒级物质各时期初始浓度图表、网格单元各时期沉积厚度表、各时期累积沉积纵横剖面图、各时期沉积平面图、各时期某网格沉积柱状图、各小层岩性百分比图表、各小层孔隙度平面等值线图等。

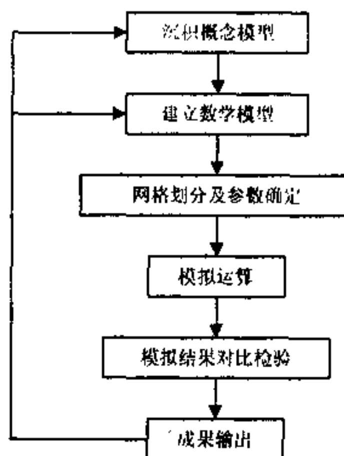


图 14-13 模拟流程图

§6 源程序及参数说明

一、剖面二维模拟源程序及参数说明

主要功能：模拟一个剖面上的沉积过程。

1. 主程序

功能：最多按 5 种粒度级完成参数输入、模拟运算、结果存储。

```
      real subsid(30),sedin(5),sedinp(5),equil(5),con(5),a(5)
      character * 14 fm1,title * 18
      common depth(30),sed(30,30,5),title,ncols,nfract
      open(4,file='sed.dat',access='direct',recl=20)
      open(5,file='depth.dat',access='direct',recl=10)
      write(*,'(a))') Enter Data filename: '
      read(*,'(a)') fm1
      open(1,file=fm1)
      read(1,*) title
      read(1,*) ntim,ncols,nfract,kprint,nplot,kplot,lag,subfac
      if(lag.lt.1) lag=1
      read(1,*) (sedinp(l),l=1,nfract)
      read(1,*) (equil(l),l=1,nfract)
      read(1,*) (con(l),l=1,nfract)
      read(1,*) (depth(l),l=1,ncols)
      read(1,*) (subsid(l),l=1,ncols)
      call crosec(0,kplot)
222   format(1x,a,i4,10i3)
      do 20 nt=1,ntim
      do 20 i=1,ncols
      do 20 l=1,nfract
      sed(nt,i,l)=0.0
```

```
20          continue
            do 110 nt=1,ntim
            if(mod(nt,kprint).eq.0) write(*,222) title,nt,(l,l=1,nfract)
            do 40 l=1,nfract
40          sedin(l)=sedinp(l)
            do 80 i=1,ncols
            do 70 l=nfract,1,-1
            a(l)=sedin(l)*con(l)
            b=depth(i)-equil(l)
            depos=amin0(a(l),b)
            if(depos.lt.0.0) depos=0.0
            sed(nt,i,l)=sed(nt,i,l)+depos
            depth(i)=depth(i)-depos
            sedin(l)=sedin(l)-depos
            if(sedin(l).lt.0.0) sedin(l)=0.0
            if(a(l).gt.b) go to 75
70          ontinue
75          is=(nt-1)*ncols+i
            write(4,rec=is) (sed(nt,i,l),l=1,nfract)
            write(5,rec=is) depth(i)
80          if(mod(nt,kprint).eq.0) write(*,223) i,depth(i),(scd(nt,i,l),
         #  sedin(l),l=1,nfract)
223         format(1x,i2,f7.2,10f7.2)
            if(mod(nt,nplot).lt.1) call crosec(nt,kplot)
            if(mod(nt,lag).gt.0) go to 110
            do 100 i=1,ncols
            sum=0.0
            if(subfac.lt.0.00001) go to 100
            do 90 lg=1,lag
            index=nt-lg+1
            do 90 l=1,nfract
90          sum=sum+sed(index,i,l)
            sum=sum*subfac
100         depth(i)=depth(i)+subsid(i)+sum
110         continue
            end
```

2. 子程序 crosec

功能:在打印机上以字符方式打印沉积剖面图。

```
            subroutine crosec(nt,kplot)
```

```
      character plot(120),plot1(120),symbol(5),title * 18
      character * 1 dot,eye,blank,rlt
      common depth(30),sed(30,30,5),title,ncols,nfract
      data symbol/'o','s','a','b','c'/
      data dot,eye,blank,rlt/' * ','|',' ','<'/
1     format(1x,a/5x,'time increment',i5)
2     format(1x,120a1)
3     format('0')
      write( * ,1) title,nt
      do 80 ii=1,ncols
      i=ncols-ii+1
      do 10 k=1,120
      plot1(k)=blank
10    plot(k)=blank
      ndep=abs(depth(i))+0.5
      index=0
      if(depth(i).lt.0.) index=ndep
      dev=abs(depth(i))-ndep
      if(ndep.lt.1) go to 30
      do 20 k=1,ndep
      index=index+1
20    if(index.ge.1) plot(index)=eye
30    if(nt.eq.0) go to 55
      do 50 nn=1,nt
      n=nt-nn+1
      sum=0.0
      lbig=0
      big=0.0
      do 35 l=1,nfract
      if(big.gt.sed(n,i,l)) go to 35
      big=sed(n,i,l)
      lbig=l
35    sum=sum+sed(n,i,l)
      nsum=sum+dev+0.5
      dev=sum+dev-nsum
      if(nsum.lt.1) go to 50
      do 40 k=1,nsum
      index=index+1
      if(index.lt.1) go to 40
      plot(index)=symbol(lbig)
```

```
        if(k.eq.nsum.and.mod(n,kplot).eq.0) plot(index)=dot
40      continue
50      continue
55      do 60 k=1,2
        index=index+1
60      if(index.ge.1) plot(index)=rlt
        if(index.lt.1) go to 70
        write(*,2) (plot(k),k=1,index)
        write(*,2) (plot1(k),k=1,index)
        go to 80
70      write(*,3)
80      continue
        end
```

3. *参数说明*

(1) 主程序

title——字符变量，存放数据文件中的标题；

ntim——整型变量，存放模拟时间阶段总数；

ncols——整型变量，存放剖面上矩形沉积单元的个数；

nfract——整型变量，存放物质粒度级数目；

nt——循环控制变量，用于从 1～ntime 的时间阶段循环；

kprint——整型变量，用于控制间隔几个时间阶段打印信息，当 nt 被 kprint 整除时打印；

nplot——整型变量，用于控制间隔几个时间阶段打印剖面图，当 nt 被 nplot 整除时打印；

kplot——整型变量，用于控制打印剖面图时的时间界面；

lag——整型变量，用于控制基底的沉降间隔；

subfac——变量，和上覆沉积物有关的沉降系数；

sedinp——数组，存放各粒度级物质的初始浓度；

equil——数组，存放各粒度级物质的基准面深度；

con——数组，存放各粒度级物质的沉积系数；

depth——数组，存放各矩形单元柱的初始水深；

subsid——矩形单元柱随时间的沉降量；

sed——三维数组，存放某阶段、某矩形柱、某粒度级物质的沉积量；

(2) 子程序

plot——字符数组，存放每次打印一行时的一组符号；

symbol ——字符数组，依次存放字符 'o'、's'、'a'、'b'、'c'，表示 5 种粒度级物质；

dot,eye,blank,rlt——字符变量，分别存放 '*'、'|'、' '、'<'，供打印使用。

二、平面二维模拟源程序及参数说明

1. 程序 1

主要功能：模拟参数输入、模拟计算、模拟结果存储。

(1) 主程序

功能：最多按5种粒度级、模拟网格30×30，进行沉积过程模拟，包括模拟参数输入、计算各阶段水深变化、物质浓度及沉积厚度、盆地沉降量等，最后存储模拟结果。

```
$debug
        dimension depth(30,30),val(30),label(30),subsid(30,30)
        dimension equil(5),con(30),ssed(5),coefx(30),coefy(30)
        character * 14 fm1
        write(*,'(a))') ' Enter Modeling Paremeter filename: '
        read(*,'(a)') fm1
        open(1,file=fm1)
        open(2,file='val.dat',access='direct',recl=8)
        open(3,file='label.dat',access='direct',recl=8)
        open(4,file='sed.dat',access='direct',recl=8)
        open(5,file='depth.dat',access='direct',recl=8)
        open(6,file='valn.dat',access='direct',recl=8)
        open(7,file='cjxs.dat',access='direct',recl=8)
        open(8,file='ksxsx.dat',access='direct',recl=8)
        open(9,file='ksxsy.dat',access='direct',recl=8)
        open(10,file='valo.dat',access='direct',recl=8)
        read(1,*)
        read(1,*) ntim,ntimd,nrows,ncols,nfract,lag,subfac,dxy,dt
        if(lag.lt.1) lag=1
        read(1,*)
        read(1,*) (equil(l),l=1,nfract)
        do 10 l=1,nfract
        read(1,*)
        do 10 i=1,nrows
        read(1,*) (con(j),j=1,ncols)
        do 10 j=1,ncols
        ic=(i-1)*ncols*nfract+(j-1)*nfract+l
        write(7,rec=ic) con(j)
10      continue
        do 20 l=1,nfract
        read(1,*)
        do 20 i=1,nrows
        read(1,*) (coefx(j),j=1,ncols)
        do 20 j=1,ncols
        ic=(i-1)*ncols*nfract+(j-1)*nfract+l
        write(8,rec=ic) coefx(j)
20      continue
        do 30 l=1,nfract
```

```
            read(1, * )
            do 30 i=1,nrows
            read(1, * ) (coefy(j),j=1,ncols)
            do 30 j=1,ncols
            ic=(i-1) * ncols * nfract+(j-1) * nfract+1
            write(9,rec=ic) coefy(j)
30          continue
            read(1, * )
            do 40 i=1,nrows
            read(1, * ) (depth(i,j),j=1,ncols)
40          continue
            read(1, * )
            do 45 i=1,nrows
            read(1, * ) (subsid(i,j),j=1,ncols)
45          continue
            do 50 l=1,nfract
            read(1, * )
            do 50 i=1,nrows
            read(1, * ) (val(j),j=1,ncols)
            do 50 j=1,ncols
            iv=(i-1) * ncols * nfract+(j-1) * nfract+l
            write(2,rec=iv) val(j)
            write(10,rec=iv) val(j)
50          continue
            read(1, * )
            do 55 i=1,nrows
            read(1, * ) (label(j),j=1,ncols)
            do 55 j=1,ncols
            il=(i-1) * ncols+j
            write(3,rec=il) label(j)
55          continue
            do 110 nt=1,ntim
            print * ,' t—no:',nt
            if(mod(nt,ntimd).eq.1.and.nt.ne.1) then
            do 60 l=1,nfract
            do 60 i=1,nrows
            do 60 j=1,ncols
            iv=(i-1) * ncols * nfract+(j-1) * nfract+l
            read(10,rec=iv) valij
            if(valij.gt.0.0) write(2,rec=iv) valij
```

```
60          continue
            end if
            call ksnd(nt,nrows,ncols,nfract,dxy,dt)
            do 80 i=1,nrows
            do 80 j=1,ncols
            do 65 l=1,nfract
            ssed(l)=0.0
65          continue
            id=(nt-1)*nrows*ncols+(i-1)*ncols+j
            do 70 l=nfract,1,-1
            iv=(i-1)*ncols*nfract+(j-1)*nfract+l
            read(2,rec=iv) valij
            read(7,rec=iv) conij
            a=valij*conij
            b=depth(i,j)-equil(l)
            depos=amin0(a,b)
            if(depos.lt.0.0) depos=0.0
            ssed(l)=ssed(l)+depos
            depth(i,j)=depth(i,j)-depos
            valij=valij-depos
            if(valij.lt.0.0) valij=0.0
            write(2,rec=iv) valij
            is=(nt-1)*nrows*ncols*nfract+(i-1)*ncols
     #      *nfract+(j-1)*nfract+l
            write(4,rec=is) ssed(l)
            write(5,rec=id) depth(i,j)
70          continue
80          continue
            if(mod(nt,lag).gt.0) go to 110
            do 100 i=1,nrows
            do 100 j=1,ncols
            sum=0.0
            if(subfac.le.0.0) go to 95
            do 90 lg=1,lag
            ind=nt-lg+1
            do 90 l=1,nfract
            is=(ind-1)*nrows*ncols*nfract+(i-1)*ncols
     #      *nfract+(j-1)*nfract+l
            read(1,rec=is) ssed(l)
90          sum=sum+ssed(l)
```

```
          sum=sum * subfac
95        subsum=subsid(i,j)+sum
          if(subsum.gt.0.0) then
          id=(nt-1) * nrows * ncols+(i-1) * ncols+j
          depth(i,j)=depth(i,j)+subsum
          write(5,rec=id) depth(i,j)
          end if
100       continue
110       continue
          end
```

(2) 子程序 ksnd

功能:计算各粒度级物质的"扩散"浓度,并将结果存入文件。

```
          subroutine ksnd(nt,nrows,ncols,nfract,dxy,dt)
          dimension val(30,30),valnew(30,30)
          do 240 l=1,nfract
          do 20 i=1,nrows
          do 20 j=1,ncols
          iv=(i-1) * ncols * nfract+(j-1) * nfract+l
          read(2,rec=iv) val(i,j)
20        continue
          do 200 i=1,nrows
          do 200 j=1,ncols
          il=(i-1) * ncols+j
          read(3,rec=il) labij
          if=(i-1) * ncols * nfract+(j-1) * nfract+l
          read(8,rec=if) coefx
          read(9,rec=if) coefy
          factx=coefx * dt/(dxy * dxy)
          facty=coefy * dt/(dxy * dxy)
          if(labij.lt.1.or.labij.gt.9) then
          write( * ,'(1x,a,2i3)') ' Data error in',i,j
          stop
          end if
          if(labij.eq.1) then
          valnew(i,j)=val(i,j)+factx * (val(i,j+1)-val(i,j))+
     #    facty * (val(i+1,j)-val(i,j))
          end if
          if(labij.eq.2) then
          valnew(i,j)=val(i,j)+factx * (val(i,j-1)+val(i,j+1)-2.0 * val(i,j))
     #    +facty * (val(i+1,j)-val(i,j))
```

```
        end if
        if(labij.eq.3) then
        valnew(i,j)=val(i,j)+factx*(val(i,j-1)-val(i,j))
     #  +facty*(val(i+1,j)-val(i,j))
        end if
        if(labij.eq.4) then
        valnew(i,j)=val(i,j)+factx*(val(i,j-1)-val(i,j))
     #  +facty*(val(i-1,j)+val(i+1,j)-2.0*val(i,j))
        end if
        if(labij.eq.5) then
        valnew(i,j)=val(i,j)+factx*(val(i,j-1)-val(i,j))
     #  +facty*(val(i-1,j)-val(i,j))
        end if
        if(labij.eq.6) then
        valnew(i,j)=val(i,j)+factx*(val(i,j-1)+val(i,j+1)-2.0*val(i,j))
     #  +facty*(val(i-1,j)-val(i,j))
        end if
        if(labij.eq.7) then
        valnew(i,j)=val(i,j)+factx*(val(i,j+1)-val(i,j))
     #  +facty*(val(i-1,j)-val(i,j))
        end if
        if(labij.eq.8) then
        valnew(i,j)=val(i,j)+factx*(val(i,j+1)-val(i,j))
     #  +facty*(val(i+1,j)+val(i-1,j)-2.0*val(i,j))
        end if
        if(labij.eq.9) then
        valnew(i,j)=val(i,j)+factx*(val(i,j-1)+val(i,j+1)-2.0*val(i,j))
     #  +facty*(val(i+1,j)+val(i-1,j)-2.0*val(i,j))
        end if
        iv=(nt-1)*nrows*ncols*nfract+(i-1)*ncols
     #  *nfract+(j-1)*nfract+1
        write(6,rec=iv) valnew(i,j)
200     continue
        do 220 i=1,nrows
        do 220 j=1,ncols
        iv=(i-1)*ncols*nfract+(j-1)*nfract+1
        write(2,rec=iv) valnew(i,j)
220     continue
240     continue
        end
```

(3) 参数说明

① 主程序

ntim——整型变量,存放模拟时间阶段总数;

ntimd——整型变量,存放时间阶段间隔数,每隔 ntimd,重新补充物源;

nrows——整型变量,存放模拟网格的行数;

ncols——整型变量,存放模拟网格的列数;

nfract——整型变量,存放物质粒度级数目;

nt——循环控制变量,用于从 1~ntime 的时间阶段循环;

lag——整型变量,用于控制基底的沉降间隔;

subfac——变量,和上覆沉积物有关的沉降系数;

dxy——变量,存放正方形网格单元的边长;

dt——变量,在一个时间阶段内又按 dt 分为若干小的时间阶段,供求解微分方程使用;

val——数组,将各粒度级物质的初始浓度存入数据文件使用的中间数组;

equil——数组,存放各粒度级物质的基准面深度;

con——数组,将各粒度级物质的沉积系数存入数据文件使用的中间数组;

depth——数组,存放各正方形网格单元的初始水深;

subsid——各正方形网格单元内随时间的沉降量;

ssed——数组,中间过程使用,存放各粒度级物质的沉积量;

lable——数组,将各正方形网格单元的类型存入数据文件使用的中间数组;

coefx——数组,将各粒度级物质在各单元的 x 方向扩散系数存入文件使用的中间数组;

coefy——数组,将各粒度级物质在各单元的 y 方向扩散系数存入文件使用的中间数组;

数据文件(随机)——

label.dat——存放各正方形网格单元的类型;

sed.dat——存放各阶段、各单元、各粒度级物质的沉积量;

depth.dat——存放各正方形网格单元在各阶段的水深;

cjxs.dat——存放各粒度级物质在各单元的沉积系数;

ksxsx.dat——存放各粒度级物质在各单元的 x 方向的扩散系数;

ksxsy.dat——存放各粒度级物质在各单元的 y 方向的扩散系数;

val.dat——存放各阶段、各单元、各粒度级物质的浓度;

valo.dat——存放各阶段、各单元、各粒度级物质的初始浓度。

② 子程序 ksnd

nt——形式参数,传递时间阶段号;

nrows——形式参数,传递模拟网格的行数;

ncols——形式参数,传递模拟网格的列数;

nfract——形式参数,传递物质的粒度级总数;

dxy——形式参数,传递每个正方形网格单元的边长;

dt——形式参数,传递将一个时间阶段内分为若干小的时间阶段的时间间隔。

2. 程序 2

主要功能:绘制某时刻的沉积纵剖面图

(1) 主程序

功能：调用由程序1保存的计算结果文件，计算一个纵剖面上各沉积单元、各粒度级物质的沉积厚度，绘出剖面上的沉积形态。

```
          real sed(5)
          open(4,file='sed.dat',access='direct',recl=8)
          open(5,file='depth.dat',access='direct',recl=8)
          write(*,'(a\)') ' Enter nt,row:'
          read(*,*) nt,ni
          write(*,'(a\)') ' Enter start—end col: '
          read(*,*) nscol,necol
          write(*,'(a\)') ' Enter nfract,nrows,ncols: '
          read(*,*) nfract,nrows,ncols
          call in
          call fact(10.0)
          dx=3.0
          dy=1.0
          print *,'t= ',nt
          do 80 j=nscol,necol
          yy=20.0*dy
          is=(nt-1)*nrows*ncols+(ni-1)*ncols+j
          read(5,rec=is) depth
          xx=j*dx
          yy=yy-depth*dy
          call yx(xx,yy,dx,depth*dy,0.2,0.2,10,-1)
          do 20 it=nt,1,-1
          do 5 l=1,nfract
          is1=(it-1)*nrows*ncols*nfract+(ni-1)*ncols
     #    *nfract+(j-1)*nfract+l
          read(4,rec=is1) sed(l)
5          continue
           do 10 l=1,nfract
           if(sed(l).le.0.0) go to 10
           yy=yy-sed(l)*dy
           ipen=l
           if(l.eq.1) ipen=10
           if(l.eq.2) ipen=12
           if(l.eq.3) ipen=15
           call yx(xx,yy,dx,sed(l)*dy,0.2,0.2,ipen,l)
10         continue
20         continue
           call pen(12)
```

```
      if(j.eq.nscol) then
      call movea(xx,yy)
      call linea(xx+dx,yy)
      else
      call movea(xxx+dx,yyy)
      call linea(xx,yy)
      call linea(xx+dx,yy)
      end if
      xxx=xx
      yyy=yy
80    continue
      end
```

(2) 子程序 yx

功能:在一个矩形框内绘制表示 5 种粒度级之一物质的符号。

```
      subroutine yx(x0,y0,w,h,dxx,dyy,ipen,l)
      call pen(ipen)
      call box(x0,y0,w,h)
      if(l.lt.1.or.l.gt.5) return
      dx=dxx
      dy=dyy
      nx=w/dx
      ny=h/dy
      if(h.le.2.0*dy) then
      ny=1
      dy=0.5*h
      end if
      do 20 j=1,ny
      yy=y0+j*dy
      if(abs(yy-y0-h).lt.0.5*dy) go to 20
      ddx=0
      nend=nx-1
      if(mod(j,2).eq.0) then
      ddx=0.5*dx
      nend=nx-2
      end if
      do 10 i=1,nend
      xx=x0+i*dx+ddx
      if(l.eq.1) then
      if(mod(i,2).ne.0) call movea(xx,yy)
      if(mod(i,2).eq.0) call linea(xx,yy)
```

```
      end if
      if(l.eq.2) then
      if(mod(i,2).eq.0) then
      call movea(xx-0.5*dx,yy)
      call linea(xx+0.5*dx,yy)
      else
      call cir(xx,yy,0.01,3)
      end if
      end if
      if(l.eq.3) call cir(xx,yy,0.01,3)
      if(l.eq.4) then
      if(mod(i,2).eq.0) call cir(xx,yy,0.01,3)
      if(mod(i,2).ne.0) call cir(xx,yy,0.08,4)
      end if
      if(l.eq.5) call cir(xx,yy,0.08,4)
10    continue
20    continue
      end
```

(3) 其它子程序

功能:绘矩形框、圆等。(见《计算机绘制地质图》,李汉林、赵永军,石油大学出版社,1997.2)

```
      subroutine box(x0,y0,w,h)
      call movea(x0,y0)
      call linea(x0+w,y0)
      call linea(x0+w,y0+h)
      call linea(x0,y0+h)
      call linea(x0,y0)
      end

      subroutine cir(x0,y0,r,n)
      dt=2.0*3.1415926/n
      call movea(x0+r,y0)
      do 5 i=2,n+1
      t=(i-1)*dt
      x=x0+r*cos(t)
      y=y0+r*sin(t)
      call linea(x,y)
5     continue
      end
```

（4）参数说明

① 主程序

nt:整型变量,存放一个键盘输入的时间阶段编号。

nj :整型变量,存放一个键盘输入的模拟网格行编号。

nscol:整型变量,存放一个键盘输入的模拟网格起始列编号。

necol:整型变量,存放一个键盘输入的模拟网格终止列编号。

其余参数和数据文件参见程序1。

② 子程序 yx

x0、y0:形式参数,传递矩形的左下角坐标。

w、h:形式参数,传递矩形的宽度和高度。

dxx、dyy:形式参数,传递对矩形的缩放比例。

ipen:形式参数,传递绘图使用的笔(颜色)号。

l:形式参数,传递物质的粒度级编号。

3. 程序3

主要功能:绘制某时刻的沉积横剖面图

（1）主程序

功能:调用由程序1保存的计算结果文件,计算一个横剖面上各沉积单元、各粒度级物质的沉积厚度,绘出剖面上的沉积形态。

```
real sed(5)
open(4,file='sed.dat',access='direct',recl=8)
open(5,file='depth.dat',access='direct',recl=8)
write(*,'(a))') ' Enter nt,col:'
read(*,*) nt,nj
write(*,'(a))') ' Enter start—end row: '
read(*,*) nsrow,nerow
write(*,'(a))') ' Enter nfract,nrows,ncols: '
read(*,*) nfract,nrows,ncols
call in
call fact(10.0)
dx=3.0
dy=1.0
print *,'t= ',nt
do 80 i=nsrow,nerow
yy=20.0*dy
is=(nt-1)*nrows*ncols+(i-1)*ncols+nj
read(5,rec=is) depth
xx=i*dx
yy=yy-depth*dy
call yx(xx,yy,dx,depth*dy,0.2,0.2,10,-1)
do 20 it=nt,1,-1
```

```
          do 5 l=1,nfract
          is1=(it-1)*nrows*ncols*nfract+(i-1)*ncols
     #    *nfract+(nj-1)*nfract+l
          read(4,rec=is1) sed(l)
5         continue
          do 10 l=1,nfract
          if(sed(l).le.0.0) go to 10
          yy=yy+sed(l)*dy
          ipen=5
          if(l.eq.1) ipen=10
          if(l.eq.2) ipen=12
          if(l.eq.3) ipen=15
          call yx(xx,yy,dx,sed(l)*dy,0.2,0.2,ipen,l)
10        continue
20        continue
          call pen(12)
          if(i.eq.nsrow) then
          call movea(xx,yy)
          call linea(xx+dx,yy)
          else
          call movea(xxx+dx,yyy)
          call linea(xx,yy)
          call linea(xx+dx,yy)
          end if
          xxx=xx
          yyy=yy
80        continue
          end
```

(2) 子程序 yx

功能：在一个矩形框内绘制表示 5 种粒度级之一物质的符号，代码与程序 2 中相同。

(3) 其它子程序

功能：绘矩形框、圆，移笔，画线，设置绘图单位，绘图初始化，选择笔号等。代码与程序 2 中相同。（见《计算机绘制地质图》，李汉林、赵永军，石油大学出版社，1997.2）

(4) 参数说明

nt：整型变量，存放一个键盘输入的时间阶段编号。

nj：整型变量，存放一个键盘输入的模拟网格列编号。

nsrow：整型变量，存放一个键盘输入的模拟网格起始行编号。

nerow：整型变量，存放一个键盘输入的模拟网格终止行编号。

其余参数和数据文件参见程序 1 和程序 2。

4. 程序 4

主要功能:绘制某时刻的沉积俯视平面图。

(1) 主程序

功能:调用由程序 1 保存的计算结果文件,计算网格平面上各沉积单元表面物质类型(5种粒度级),绘出平面上的沉积形态。

```
              real sed(5)
              open(4,file='sed.dat',access='direct',recl=8)
              write(*,'(a))') ' Enter nt: '
              read(*,*) nt
              write(*,'(a))') ' Enter start--end row: '
              read(*,*) nsrow,nerow
              write(*,'(a))') ' Enter start--end col: '
              read(*,*) nscol,necol
              write(*,'(a))') ' Enter nfract,nrows,ncols: '
              read(*,*) nfract,nrows,ncols
              call in
              call fact(10.0)
              dxy=2.0
              print *,'t= ',nt
              do 80 i=nsrow,nerow
              yy=40.0-(i-1)*dxy
              do 80 j=nscol,necol
              xx=j*dxy
              ipen=10
              do 20 it=nt,1,-1
              do 5 l=1,nfract
              is=(it-1)*nrows*ncols*nfract+(i-1)*ncols
          #   *nfract+(j-1)*nfract+l
              read(4,rec=is) sed(l)
              ll=-1
              if(sed(l).gt.0.0) then
              ll=l
              ipen=l
              go to 25
              end if
5             continue
20            continue
25            call yx(xx-0.5*dxy,yy-0.5*dxy,dxy,dxy,0.2,0.2,ipen,ll)
80            continue
              end
```

（2）子程序 yx

功能：在一个矩形框内绘制表示 5 种粒度级之一物质的符号，代码与程序 2 中相同。

（3）其它子程序

功能：绘矩形框、圆，移笔，画线，设置绘图单位，绘图初始化，选择笔号等。代码与程序 2 中相同。（见《计算机绘制地质图》，李汉林、赵永军，石油大学出版社，1997.2）

（4）参数说明

nt：整型变量，存放一个键盘输入的时间阶段编号。

nsrow：整型变量，存放一个键盘输入的模拟网格起始行编号。

nerow：整型变量，存放一个键盘输入的模拟网格终止行编号。

nscol：整型变量，存放一个键盘输入的模拟网格起始列编号。

necol：整型变量，存放一个键盘输入的模拟网格终止列编号。

其余参数和数据文件参见程序 1 和程序 2。

5. 程序 5

主要功能：绘制某时刻、某沉积单元内的柱状剖面图。

（1）主程序

功能：调用由程序 1 保存的计算结果文件，计算网格平面一个选定沉积单元内，各粒度级物质的沉积厚度和沉积顺序，绘出柱状剖面图。

```
      real sed(5)
      open(4,file='sed.dat',access='direct',recl=8)
      open(5,file='depth.dat',access='direct',recl=8)
      write(*,'(a))') ' Enter nt,row,col:'
      read(*,*) nt,ni,nj
      write(*,'(a))') ' Enter nfract,nrows,ncols,dx,dy,ytop: '
      read(*,*) nfract,nrows,ncols,dx,dy,yy
      call in
      call fact(10.0)
      print *,'t= ',nt
      is=(nt-1)*nrows*ncols+(ni-1)*ncols+nj
      read(5,rec=is) depth
      xx=dx
      yy=yy-depth*dy
      nnn=0
      do 20 it=nt,1,-1
      do 5 l=1,nfract
      is=(it-1)*nrows*ncols*nfract+(ni-1)*ncols
     #*nfract+(nj-1)*nfract+l
      read(4,rec=is) sed(l)
5     continue
      do 10 l=1,nfract
      if(sed(l).le.0.0) go to 10
```

```
          nnn=nnn+1
          yy=yy-sed(l) * dy
          ipen=1
          if(l.eq.1) ipen=10
          if(l.eq.2) ipen=12
          if(l.eq.3) ipen=15
          call yx(xx,yy,dx,sed(l) * dy,0.2,0.2,ipen,l)
10        continue
20        continue
          print * ,' sed—times: ',nnn
          end
```

(2) 子程序 yx

功能：在一个矩形框内绘制表示 5 种粒度级之一物质的符号，代码与程序 2 中相同。

(3) 其它子程序

功能：绘矩形框、圆，移笔，画线，设置绘图单位，绘图初始化，选择笔号等。代码与程序 2 中相同。(见《计算机绘制地质图》，李汉林、赵永军，石油大学出版社，1997.2)

(4) 参数说明

nt：整型变量，存放一个键盘输入的时间阶段编号。

ni：整型变量，存放一个键盘输入的模拟网格行编号。

nj：整型变量，存放一个键盘输入的模拟网格列编号。

dx：变量，将柱状剖面图在横向扩大 dx 倍。

dy：变量，将柱状剖面图在垂向扩大 dy 倍。

yy：变量，调整柱状剖面图垂向上的位置。

其余参数和数据文件参见程序 1 和程序 2。

6. 程序 6

(1) 主程序

功能：调用由程序 1 保存的计算结果文件，计算各沉积单元内，各粒度级物质的沉积总厚度和各单元中心的相对坐标，形成绘等值线图所用的数据文件(文件名由键盘输入)。

```
$debug
          real x(30),y(30)
          character * 14 fm
          write( * ,'(a ))') 'Enter New Filename: '
          read( * ,'(a)') fm
          open(1,file=fm)
          write( * ,'(a ))') ' Enter nt,ncols,nrows,nfract,ll:'
          read( * , * ) nt,ncols,nrows,nfract,ll
          write( * ,'(a ))') ' Enter x0,y0,dxy:'
          read( * , * ) x0,y0,dxy
          do 5 i=1,ncols
          x(i)=x0+(i-1) * dxy
```

```
5          continue
           do 10 j=1,nrows
           y(j)=y0+(j-1)*dxy
10         continue
           open(4,file='sed.dat',access='direct',recl=8)
           do 80 i=1,nrows
           do 80 j=1,ncols
           ssed=0.0
           do 20 it=nt,1,-1
           is=(it-1)*nrows*ncols*nfract+(i-1)*ncols
      #    *nfract+(j-1)*nfract+ll
           read(4,rec=is) sed
           if(sed.gt.0.0) then
           ssed=ssed+sed
           end if
20         continue
           write(1,111) x(j),y(i),ssed
80         continue
           close(1)
           close(4)
111        format(1x,f8.2,',',f8.2,',',f8.4)
           end
```

(2) 参数说明

ll:整型变量,存放粒度级编号。

x0,y0:变量,存放左下角沉积单元中心坐标。

dxy:变量,存放正方形单元的边长。

其余参数和数据文件见程序 1 和程序 2。

§7 应用算例

本节提供了基于水下扇的一批模拟参数,使用上述程序进行计算并绘出了部分图件。

一、模拟参数

模拟参数在计算机上以数据文件的方式建立,数字之间用逗号或空格分隔。这里考虑了 3 种粒度的物质,粒度级由细到粗,编号从 1~3,模拟网格为 30×30。实际文件中不包含带括号的汉字部分,在此仅用于提示。参数数据文件如下:

ntim,ntimd,nrows,ncols,nfract,lag,subfac,dxy,dt (**用于提示,见程序 1 参数说明**)

24,1,20,20,3,4,0.1,1.0,0.25

equil (**粒度级 1~3 物质的基准面深度**)

5.0,3.0,1.0

cjxs size: 1 (**粒度级 1 物质的沉积系数**)

0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9

cjxs size：2（**粒度级 2 物质的沉积系数**）

0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9

0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9

cjxs size：3（粒度级3物质的沉积系数）

0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9
0.01 0.01 0.02 0.02 0.02 0.04 0.06 0.08 0.1 0.2 0.2 0.3 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.9

ksxs—x size：1（粒度级1物质在x方向的扩散系数）

0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9

0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9

ksxs—x size：2（**粒度级 2 物质在 x 方向的扩散系数**）

0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9

ksxs—x size：3（**粒度级 3 物质在 x 方向的扩散系数**）

0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9

0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9

ksxs—y size：1（**粒度级 1 物质在 y 方向的扩散系数**）

0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9

ksxs—y size：2（**粒度级 2 物质在 y 方向的扩散系数**）

0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.3 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.80.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.80.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9

0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9

ksxs—y size；3（**粒度级 3 物质在 y 方向的扩散系数**）

0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.4 0.4 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.5 0.5 0.5 0.5 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.8 0.9
0.9 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.8 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9

depth（**湖盆原始水深**）

1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
1.0 1.0 2.0 3.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 3.0 2.0 1.0 1.0
1.0 2.0 3.0 4.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 4.0 3.0 2.0 1.0
1.0 2.0 3.0 6.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 6.0 3.0 2.0 1.0
1.0 2.0 3.0 7.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 4.0 2.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 4.0 2.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0

1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 9.0 8.0 7.0 4.0 1.0
1.0 2.0 3.0 7.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 8.0 4.0 2.0 1.0
1.0 2.0 3.0 6.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 7.0 6.0 3.0 2.0 1.0
1.0 2.0 3.0 4.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 5.0 4.0 3.0 2.0 1.0
1.0 2.0 2.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 4.0 3.0 2.0 1.0 1.0
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0

subsid（**湖盆各网格单元的沉降量**）

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

val size：1（**粒度级 1 物质的初始浓度**）

0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
0 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0
660 0 0 0   0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0

660 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
660 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
660 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

val size：2（**粒度级 2 物质的初始浓度**）

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
520 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
520 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
520 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
520 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

val size：3（**粒度级 3 物质的初始浓度**）

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

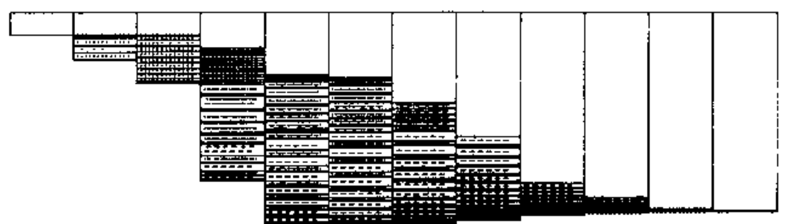
```
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
160 0 0 0    0 0 0 0 0 0 0 0 0 0     0 0 0 0 0 0
160 0 0 0    0 0 0 0 0 0 0 0 0 0     0 0 0 0 0 0
160 0 0 0    0 0 0 0 0 0 0 0 0 0     0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
0 0 0 0      0 0 0 0 0 0  0 0 0 0    0 0 0 0 0 0
```

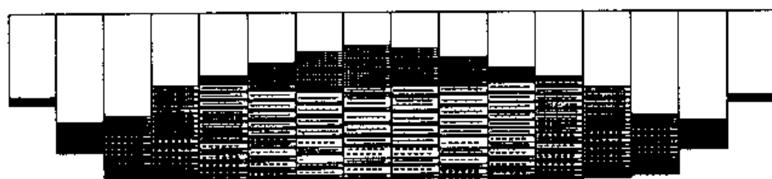
label（各网格单元类型）

```
1 2 2 2      2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
8 9 9 9      9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 4
7 6 6 6      6 6 6 6 9 9 9 9 9 9 9 9 9 9 6 5
```

二、部分模拟结果图件

模拟过程共经历 24 个时间阶段，以下列出了 nt=24 时的部分模拟图件。





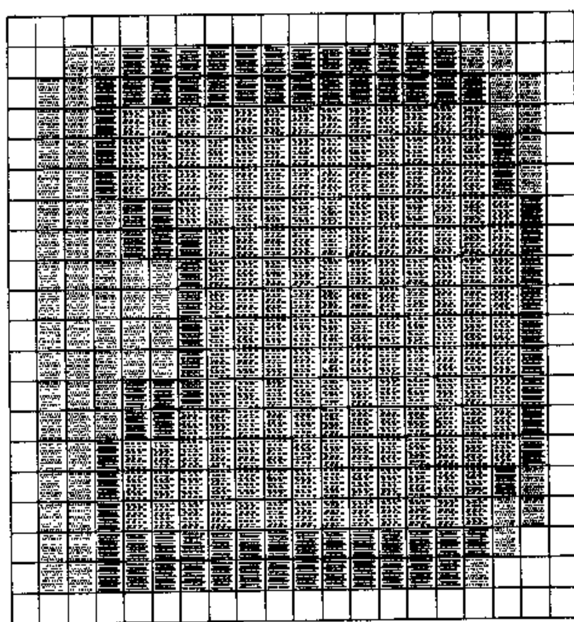


图 14-14 模拟剖面图及平面图(nt=24)

(a) 纵剖面图(11 行,1～12 列)

(b) 横剖面图(3～18 行,4 列) (c)物质平面分布图

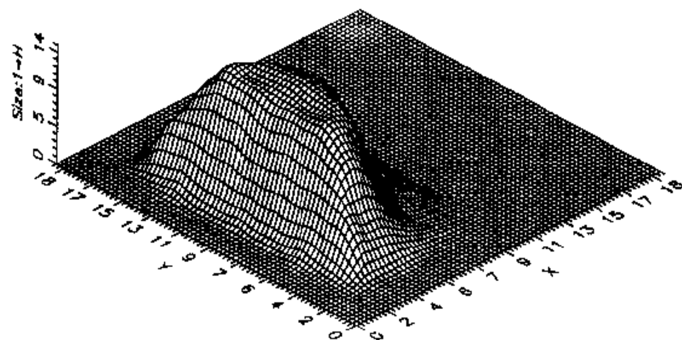


图 14-15　粒度 1 物质总厚度曲面图(nt=24)

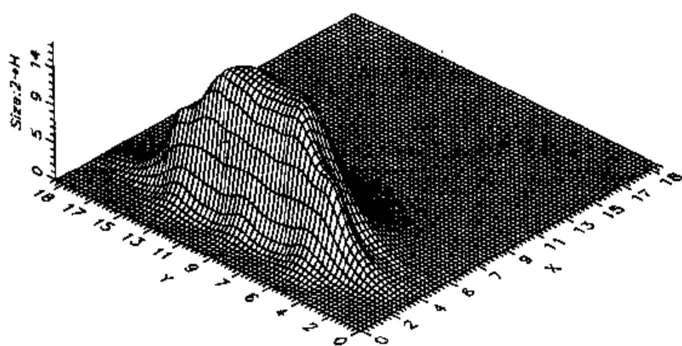


图 14-16　粒度 2 物质总厚度曲面图(nt=24)

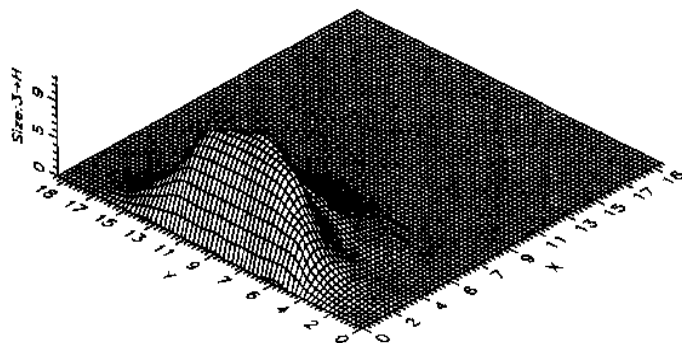


图 14-17　粒度 3 物质总厚度曲面图(nt=24)

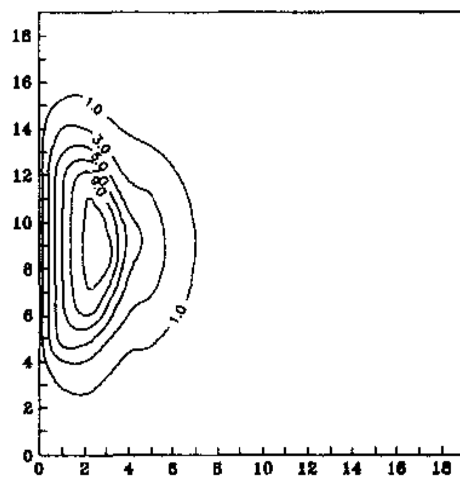


图 14-18　粒度 1 物质总厚度等值线图(nt=24)

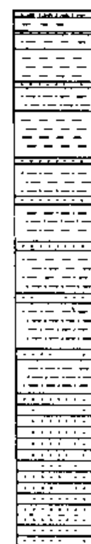


图 14-19　柱状图(nt=24,11 行,4 列)

第十五章　石油资源量及含油气有利地带的预测

油气资源评价成果是发展石油工业的基础，因此，石油资源量及含油气有利地带的合理预测就成为整个石油地质勘探过程中的重要任务之一。

本世纪70年代以来，由于电子计算机技术的迅猛发展和在石油勘探中的普遍推广、数学地质方法的应用、极大地促进了石油资源评价工作的开展，并且提高了评价方法的科学性和评价结果的可靠性。由于石油资源评价的任务、理论基础和勘探开发阶段的不同，又有不同的石油资源量及含油气有利地带的评价方法。据不完全统计，目前国内外用于石油资源评价的方法多达百余种，本书不作一一介绍，除前面介绍的蒙特卡洛法和盆地数值模拟外，再介绍几种常用的石油资源量及含油气有利地带的预测方法。

§1　石油资源量预测

一、翁(Weng)旋回预测模型

翁旋回模型是由我国著名科学家翁文波先生提出的。如果某一体系具有从兴起到衰亡的完整过程，则这一过程可称作一个生命旋回。对于生命总量有限的一些体系，例如对于非再生矿产资源(油、气)的开采，可用Weng旋回预测模型进行描述和预测。

1. Weng旋回预测模型及其性质

客观世界中“从无到有”突然出现的实际体系，是一种不连续体系，记为Q。如果将体系Q出现的时间记为0，体系Q的不连续性可以表示为：

$$Q=\begin{cases}0 & (t<0)\\ Q & (t>0)\end{cases}$$

如果体系的发展速度$\mathrm{d}Q/\mathrm{d}t$正比于实际存在的“现状”或“基础”Q，那么为了引入不连续过程，假设$\mathrm{d}Q/\mathrm{d}t$正比于因子$(x/t-1)$，其中x为Q达到顶峰的时间($x>0$)，因此有：

$$\frac{\mathrm{d}Q}{\mathrm{d}t}=Q\left(\frac{x}{t}-1\right)$$

解上面的微分方程得到：

$$\ln Q=x\ln t-t+\ln A$$

上式中$\ln A$为积分常量。若认为$t=0$时，$Q=0$，则整理上式可得：

$$Q=At^{x}e^{-t}\qquad(t\geqslant 0)\tag{15-1}$$

式(15-1)就是Weng旋回预测模型。从该式可看出，体系Q的兴衰正比于兴起和衰亡两个因子。Q的兴起正比于时间t的x次方(兴起因子)，Q的衰亡正比于时间t的负指数函数(衰亡因子)。因此，体系Q是时间t的函数，而t是时间间隔与系数C的比值。所以，Weng旋回预测模型又可以表示为：

$$\begin{cases}Q_t=At^{x}e^{-t}\\ t=(T-T_0)/C\end{cases}\qquad(t\geqslant 0)\tag{15-2}$$

式中　x——某一正实数；

T_0——生命起始时刻；

T——生命过程中的某时刻；

A、C——拟合系数。

Weng 旋回预测模型具有下列性质：

① $\frac{\mathrm{d}Q_t}{\mathrm{d}t} = Axt^{x-1}e^{-1} - At^x e^{-1} = Ax\frac{t^x}{t}e^{-1} - At^x e^{-1} = At^x e^{-t}(\frac{x}{t} - 1)$

即 $\frac{\mathrm{d}Q_t}{\mathrm{d}t} = Q_t(\frac{x}{t} - 1)$

所以：当 $t<x$ 时，$\frac{\mathrm{d}Q}{\mathrm{d}t}>0$，表示体系 Q 兴起。

当 $t=x$ 时，$\frac{\mathrm{d}Q}{\mathrm{d}t}=0$，表示体系 Q 达到高峰。

当 $t>x$ 时，$\frac{\mathrm{d}Q}{\mathrm{d}t}<0$，表示体系 Q 衰亡。

②
$$\begin{aligned}\frac{\mathrm{d}^2Q_t}{\mathrm{d}t^2} &= A[x(x-1)t^{x-2}e^{-t} - xt^{x-1}e^{-t} - xt^{x-1}e^{-t} + e^{-t}t^x] \\ &= At^x e^{-t}\left[\frac{x(x-1)}{t^2} - \frac{2x}{t} + 1\right] \\ &= Q_t\left[\frac{x(x-1) - 2xt + t^2}{t^2}\right] \\ &= Q_t\left[\frac{1}{t^2}(x^2 - x - 2xt + t^2)\right]\end{aligned}$$

即 $$\frac{\mathrm{d}^2Q_t}{\mathrm{d}t^2} = Q_t\frac{1}{t^2}[(t-x)^2 - x)]$$

所以：当 $t=x+\sqrt{x}$ 时，$\frac{\mathrm{d}^2Q_t}{\mathrm{d}t^2} = 0$ 。

当 $t=x-\sqrt{x}$ 时，$\frac{\mathrm{d}^2Q_t}{\mathrm{d}t^2} = 0$。

③ 当式(15-2)中的 x 为正整数，$t\to\infty$时，对 Q_t 积分可得：

$$\int_0^\infty Q_t\mathrm{d}t = A\Gamma(x+1) = Ax! = \sum_\infty Q_t$$

$\sum_\infty Q_t$ 可称作体系 Q 的生命总量。

④ $$\frac{Q_t}{\sum_\infty Q_t} = \frac{t^x e^{-t}}{x!}$$

上述表达式和单项泊松分布在形式上相同。

⑤ 确定型的 Weng 旋回，体系 Q 截至时间 t 的生命量可记为 $\sum_t Q_t$ 。如果 x 是正整数 0，1，2，…，则可导出：

$$\begin{aligned}\frac{\sum_t Q_t}{\sum_\infty Q_t} &= \frac{\int_0^t Q_t\mathrm{d}t}{\int_0^\infty Q_t\mathrm{d}t} = \frac{1}{x!}\int_0^t t^x e^{-t}\mathrm{d}t = -\frac{1}{x!}\int_0^t t^x \mathrm{d}e^{-t} \\ &= -\frac{t^x}{x!}e^{-t} + \frac{1}{x!}\int_0^t xt^{x-1}e^{-t}\mathrm{d}t \\ &= -\frac{t^x}{x!}e^{-t} - \frac{t^{x-1}}{(x-1)!}e^{-t} - \cdots - \frac{t^2}{2!}e^{-t} + \int_0^t e^{-t}\mathrm{d}t \\ &= -\frac{t^x}{x!}e^{-t} - \frac{t^{x-1}}{(x-1)!}e^{-t} - \cdots - \frac{t^2}{2!}e^{-t} - e^{-t} + 1\end{aligned}$$

$$= 1 - e^{-t}\sum_{i=0}^{x}\frac{t^i}{i!} \tag{15-3}$$

因此，Weng 旋回预测模型是收敛模型，适用于生命总量有限体系的描述和预测。$t=0$ 时，式(15-3)等于 0。$t=\infty$时，式(15-3)等于 1。在 Q 的发展过程中，式(15-3)也是时间 t 的函数。

体系 Q 从兴起到衰亡大体可分为四个阶段，即：

① 加速上升阶段：$t=0\sim(x-\sqrt{x})$。

② 一般上升阶段：$t=(x-\sqrt{x})\sim x$。

③ 一般下降阶段：$t=x\sim(x+\sqrt{x})$。

④ 缓慢下降阶段：$t=(x+\sqrt{x})\sim\infty$。

由 Weng 旋回预测模型的性质⑤可以导出：

$$\sum\nolimits_{\infty}Q_t = \sum\nolimits_{t}Q_t/(1 - e^{-t}\sum_{i=0}^{x}\frac{t^i}{i!}) \tag{15-4}$$

对于生命总量有限体系，$\sum_t Q_t$ 的值可以通过实际观测获得，因而通过式(15-4)可以预测出 Q 的生命总量。对于非再生的石油资源，生命总量 $\sum_{\infty} Q_t$ 就是一个油田的最终可采储量。因此，式(15-4)可用于预测石油可采储量。

2. 油田产量及最终可采储量的预测

油气田的形成是石油地质历史演变的结果，油气田中的石油、天然气是有限资源。油气田一经投入开发就成为一个体系，从油气田投产到产量枯竭是一个生命旋回。

用式(15-2)预测油气田产量的未来变化及最终可采储量时，式中的 x 为 0，1，2，…等正整数。T_0 为油气田的投产年份。T 为油气田投产后的开采年份。A、C 为表示油气田地质特征及开采方式的系数。

为确定式(15-2)中的拟合系数 A，在实际计算时可作如下考虑，即当油气田的 m 个已知的逐年实际产量 $Q_i(i=1,2,\cdots,m)$ 与式(15-2)中的 t^xe^{-t} 之间的相关系数最大时，认定 x 及 C 的值为最佳值，此时可求拟合系数 A。令 Q_{t_i} 为与 Q_i 相应的预测值，则有：

$$S = \sum_{i=0}^{m}(Q_i - Q_{t_i})^2 = \sum_{i=0}^{m}(Q_i - At_i^xe^{-t_i})^2$$

再令

$$\frac{dS}{dA} = 2\sum_{i=0}^{m}(Q_i - At_i^xe^{-t_i})(-t_i^xe^{-t_i}) = 0$$

$$\sum_{i=0}^{m}[A(t_i^xe^{-t_i})^2 - Q_it_i^xe^{-t_i}] = 0$$

$$A\sum_{i=0}^{m}(t_i^xe^{-t_i})^2 - \sum_{i=0}^{m}Q_i(t_i^xe^{-t_i}) = 0$$

即

$$A = \frac{\sum_{i=0}^{m}Q_i(t_i^xe^{-t_i})}{\sum_{i=0}^{m}(t_i^xe^{-t_i})^2} \tag{15-5}$$

$$R=\frac{\sum_{i=0}^{m}[(t_i^x e^{-t_i})-\overline{t_i^x e^{-t_i}}](Q_i-\overline{Q}_i)}{\sqrt{\sum_{i=0}^{m}[(t_i^x e^{-t_i})-\overline{t_i^x e^{-t_i}}]^2\sum_{i=0}^{m}(Q_i-\overline{Q}_i)^2}} \tag{15-6}$$

式(15-6)中：

$$\overline{Q}_i=\frac{1}{m}\sum_{i=0}^{m}Q_i,\quad \overline{t_i^x e^{-t_i}}=\frac{1}{m}\sum_{i=0}^{m}(t_i^x e^{-t_i})$$

因此，可用迭代法求出拟合系数 x、C。确定拟合系数 x、C 时，除了要考虑相关系数 R 尽可能大以外，还要使 Q_t 与最近时期的油田实际产量 $Q_i(i=m,m-1,\cdots)$ 尽可能接近。对于非正规开采的油田尤其如此。也就是说，在拟合时要尽量考虑近期产量，而早期产量可较少考虑。

在预测天然气田的产量时，特别是预测一个大的天然气区或一个国家乃至全球的天然气产量时，Weng 旋回预测模型中应增加一个常数项 Q_0，即：

$$\begin{cases}Q_t=Q_0+At^x e^{-t}\\ t=(T-T_0)/C\end{cases}\qquad (t\geqslant 0) \tag{15-7}$$

上式中的常数项 Q_0 可能包括目前开采工艺水平下尚不能完全采出的非正规天然气、地下水中的部分溶解气，也可能包括还在继续生成的生物气。可见常数项 Q_0 为发散部分，因而式(15-7)已与生命总量有限体系的含义有出入。但 Q_0 值一般不大，从数学上考虑，式(15-7)只是把式(15-2)从坐标原点(0,0)沿纵坐标平移一段距离，这段距离的长度等于 Q_0。因此，虽然模型中增加了一个发散部分，但并不影响模型的求解和使用。

3. 算例

北京石油勘探开发科学研究院赵旭东等人应用 Weng 旋回预测模型，曾经对国内外 170 多个油气田的年产量及最终可采储量进行过预测。计算结果表明，这些油气田的已知实际产量与 Weng 旋回预测模型的拟合值之间的相关系数绝大数都大于 0.9。而正规开发的油气田的相关系数都在 0.95 以上。

应用 Weng 旋回模型预测时，已知的实际采油年数应该大于或等于 5，也就是说，原始数据点数 $m\geqslant 5$。

罗马什金油田是苏联仅次于萨马特洛尔油田的第二大油田，位于鞑靼自治共和国东部，发现于 1948 年，1952 年投入开发。储油层为泥盆系 $Д_1$ 及 $Д_0$ 层砂岩，油层有效厚度 15 m，埋藏深度 1650～1850 m；油田面积 3800 km²，地质储量 45×10^8t。设计采收率为 53.1%，可采储量 24×10^8t。油层孔隙度为 15%～20%，平均渗透率 300～400 mD($1\text{D}=0.987\times10^{-12}\text{m}^2$)。原始地层压力为 175 atm。1956 年至 1974 年期间的产量在苏联占第一位。1970 年为产量高峰年，其年产量为 8150×10^4t，8000×10^4t 的年产量保持了 6 年。稳产期末综合含水率为 47.2%，累计采油 11.967×10^8t，采出程度为 26.59%(采出可采储量的 49.86%)。1976 年油田进入下降阶段，1976 年至 1979 年期间每年产量递减 225×10^4～430×10^4t，年递减率为 2.8%～5.9%，至 1979 年底已累计采油 14.898×10^8t，采出程度为 33.11%，含水率在 60%以上。

经过计算得出罗马什金油田的 Weng 旋回预测模型表达式为：

$$\begin{cases}Q_t=6002.3\,t^3e^{-t}\\ t=(T-1951)/6.78\end{cases}$$

按 Weng 旋回模型预测的罗马什金油田的最终可采储量 $\sum_{\infty}Q_t=25\times10^8\text{t}$ 。1952 年至

1979 年期间的已知实际年产量与 Weng 旋回模型预测值之间的相关系数为 0.99。

罗马什金油田的实际年量以及用 Weng 旋回模型预测的年产量见表 15-1 及图 15-1。

表 15-1 罗马什金油田的产量预测表

年份	实际年产量/10^4t	预测年产量/10^4t	年份	实际年产量/10^4t	预测年产量/10^4t
1952	200	16.6	1972	8000	8056.1
1953	300	114.7	1973	8000	7992.4
1954	500	334.1	1974	8000	7880.2
1955	1000	683.3	1975	8000	7725.6
1956	1400	1151.5	1976	7775	7534.7
1957	1900	1716.9	1977	7500	7313.2
1958	2400	2352.5	1978	7230	7066.8
1959	3050	3030.1	1979	6800	6800.7
1960	3800	3722.7	1980		6519.6
1961	4400	4406.3	1981		6227.8
1962	5000	5060.5	1982		5929.2
1963	5600	5669.0	1983		5627.4
1964	6040	6219.2	1984		5325.3
1965	6600	6702.5	1985		5025.6
1966	6800	7113.3	1986		4730.4
1967	7000	7449.1	1987		4441.7
1968	7600	7709.6	1988		4160.9
1969	7900	7896.8	1989		3889.4
1970	8150	8013.8	1990		3628.0
1971	8000	8065.1			

二、油田规模序列法

“油田规模”(Oilfield Size)是指油气田的最终可采储量。如果某个含油气区经过详细勘探后，发现了全部油气田，并且查明了每个油气田的最终可采储量，那么，按最终可采储量由大到小进行排列，所得的顺序称为油田的规模序列。

1. 油田规模序列法的内涵

国内外许多含油气区的统计资料表明，当一个含油气区的一些油气田被发现后，如果以油田规模为纵坐标，以油田规模的序号为横坐标，在双对数坐标纸上展点作图大致可以得到一条直线，见图 15-2。

根据这一规律，可以在探区的早期或中期勘探阶段，由已发现油气田的油田规模序列，预测尚未发现的油气田储量以及整个探区的油气总储

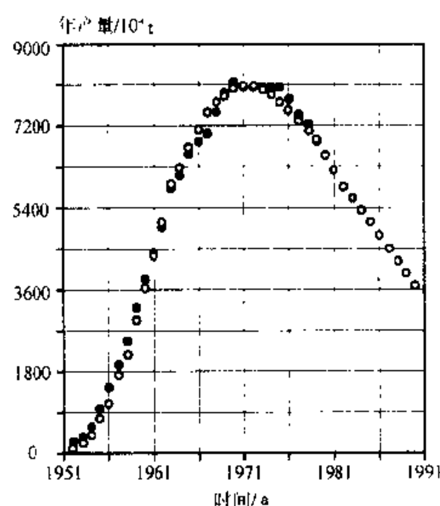


图 15-1 罗马什金油田的产量预测图

●实测 ○预测

量。这种预测方法称为油田规模序列法。

齐波夫(G.P.Zipf)于1949年在他所著的《人类行为与最小省力原则》一书中提出一种规律：对一组离散型随机变量，按取值由大到小进行排列，如果最大的数值是第二大数值的两倍，是第三大数值的三倍，……，依此类推，则称这组离散型随机变量服从齐波夫定律。

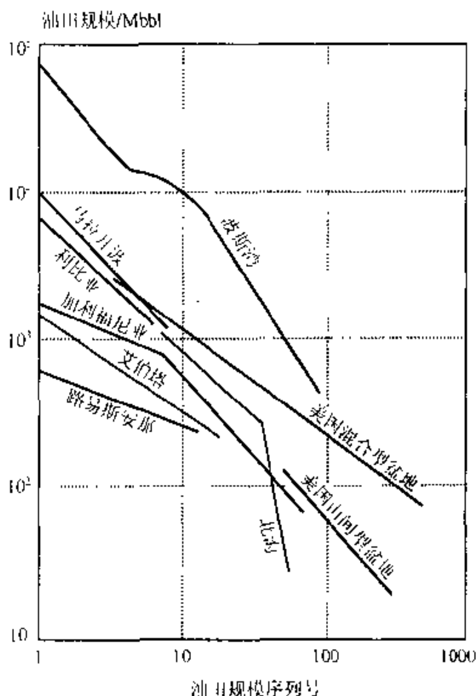


图15-2 世界主要含油气地区的油田规模序列

本世纪70年代以后，随着计算机技术的普及应用，齐波夫定律逐渐为人们所重视，1977年考兰德首先用齐波夫定律预测了赞比亚铜矿带的铜总量，其结果得到了在该地工作多年地质学家们的认可。1980年刘序琼在我国用齐波夫定律预测了一个铀矿床的资源。其后，人们应用齐波夫定律研究勘探地区的金属矿床或油气田的规模序列，借以预测尚未发现的金属矿产资源或油气资源。

实际上，齐波夫定律是巴内托(Pareto)于1927年所提出定律的特例。巴内托定律可以表述为如下关系式：

$$\frac{Q_m}{Q_n} = \left(\frac{n}{m}\right)^k \tag{15-8}$$

式中 Q_m——序号等于 m 的随机变量取值；

Q_n——序号等于 n 的随机变量取值；

k——实数；

m、n——1，2…整数序列中的任一数值，$m \neq n$。

当式(15-8)中的 $k=1$ 时，则为齐波夫定律，即：

$$\frac{Q_m}{Q_n} = \frac{n}{m} \tag{15-9}$$

或者

$$mQ_m = nQ_n$$

一个含油气地区内一组油气田的石油储量属于离散型随机变量，设最大的油田已被发现，石油储量为 Q_{max}，若油田规模序列符合齐波夫定律，则有：

$$Q_{max} = nQ_n \tag{15-10}$$

或

$$Q_n = \frac{Q_{max}}{n}$$

假如含油气区总共有 t 个油气田，则全区的石油总量储量 SQ 为：

$$SQ = \sum_{i=1}^{t}\left(\frac{Q_{max}}{i}\right) \tag{15-11}$$

对式(15-9)两边取对数，则有：

$$\lg \frac{Q_m}{Q_n} = \lg \frac{n}{m}$$

即
$$\lg Q_m - \lg Q_n = -(\lg m - \lg n)$$

即
$$\frac{\lg Q_m - \lg Q_n}{\lg m - \lg n} = -1 \quad (15\text{-}12)$$

因而，在双对数坐标纸上，以油田的石油储量 Q_i 为纵坐标，以油田的序号 i 为横坐标作图，则数据点的连线为斜率等于－1 的直线。以上的式(15-9)、(15-10)、(15-11)、(15-12)是预测矿产资源或油气资源的齐波夫定律的不同表达形式。

但是，世界上主要含油气区的多数地区并不符合齐波夫定律(图 15-2)，而是符合适应范围更广的巴内托定律。

对式(15-8)两边取对数，则有：

$$\lg \frac{Q_m}{Q_n} = \lg\left(\frac{m}{n}\right)^k$$

即
$$\frac{\lg Q_m - \lg Q_n}{\lg m - \lg n} = -k \quad (15\text{-}13)$$

因而在双对数坐标纸上作图，则数据点的边线为斜率等于－k 的直线，这样便与图 15-2 中的所有含油气地区的统计规律相符合了。所以，应当认为油田规模序列的分布规律服从巴内托定律。而齐波夫定律仅是巴内托定律的特例。

2. 油田规模序列法的使用条件

油田规模序列法的实质是根据已发现的油气田储量，应用巴内托定律预测一个含油气地区中尚未发现的油气田储量(或资源量)以及全区总的石油储量(或资源量)的一种外推预测方法。虽然世界上多数含油气地区的油田规模序列在一定程度上符合巴内托定律，然而，直至目前为止尚不能从油气形成的地质理论上圆满地解释油田规模序列的地质成因。但是，许多事实说明，任何地质过程都受概率法则支配，所以对于一个含油气地区的油田规模序列形成的原因，暂且可以从统计规律方面去理解。

油田规模序列法适用于一个完整的、独立的石油地质体系。所谓一个完整的、独立的石油地质体系是指该地质体系内的油气生成、运移、聚集以及其后的地质变迁都是在同一石油地质演化历史条件下发生的。或者说，目前所要预测的含油气地区中的油气田(或油气藏)具有统一的形成原因。

根据国内外主要含油地区的统计资料，式(15-13)中系数 k 值的变化范围在 0.5 至 2.0 之间。这一情况说明，石油地质问题的复杂性导致了油田规模序列分布的多样性。而系数 k 等于－1 的齐波夫定律只是多种油田规模序列分布的特殊情况。

当一个大的含油气地区具有多期成油过程时，可能存在多个油田的规模序列。在这种情况下需要对多个序列的复合总体进行筛分，分解出成因不同或成油期不同的多个相互独立的油田规模序列。

3. 油田规模序列法的计算过程

① 油田规模序列的系数 k 可由熟悉含油气区情况的地质家们商定。一般可以借鉴与本含油气地区在地质条件上相似的含油气地区的资料。如果确定系数 k 有困难，可令 $k = -\operatorname{tg}\theta$，而 θ 的角度值应限定在 115°～155°范围内，并把这一区间分为若干个子区间，进行多次油田规模序列的拟合计算。例如，取角度值步长为 5°时，则有如下 9 个区间间隔值：

$-\operatorname{tg} 115° = 2.1445$　　$-\operatorname{tg} 120° = 1.7321$　　$-\operatorname{tg} 125° = 1.4281$

$-\text{tg}\,130° = 1.1918$ $-\text{tg}\,135° = 1.0000$ $-\text{tg}\,140° = 0.8391$

$-\text{tg}\,145° = 0.7002$ $-\text{tg}\,150° = 0.5774$ $-\text{tg}\,155° = 0.4663$

其中 $-\text{tg}\,135° = 1$ 时为齐波夫定律。

② 把探区中已发现的 t 个油田，按储量 $Q_i(i=1,2,\cdots,t)$ 由大到小进行排列，选择最大的油田储量 Q_1 作为推算点。

③ 如果探区中已发现的 t 个油田储量为 $Q_1,Q_2,\cdots,Q_t$，则以推算点 Q_1 除 Q_i，并求出其值的 k 次方根，得到如下序列 A_i，即：

$$A_i = \sqrt[k]{\frac{Q_i}{Q_1}} \qquad (i = 1,2,\cdots,t) \tag{15-14}$$

④ 将序列 A_i 中的每个元素，乘以某一正整数 n_i，令 $A_in_i=b_i$，当所有已发现的 t 个油田的 b_i 值 $(i=1,2,\cdots,t)$ 都最大限度地接近下面矩阵某一行号 h 时记为下面的矩阵 B 的一行，依次令行号 $h=1、2、\cdots,m$，可得：

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1t} \\ b_{21} & b_{22} & \cdots & b_{2t} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m1} & b_{m2} & \cdots & b_{mt} \end{bmatrix} \begin{matrix} b_{1i} \approx 1 \\ b_{2i} \approx 2 \\ \cdots \\ b_{mi} \approx m \end{matrix}$$

计算矩阵中各行的标准差 σ_h：

$$\sigma_h = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(b_{hi} - \bar{b}_h)^2} \qquad (h = 1,2,\cdots,m)$$

$$\bar{b}_h = \frac{1}{t}\sum_{i=1}^{t} b_{hi}$$

当计算至矩阵中第 m 行的标准差 σ_m 小于给定误差 ε 时，即 $\sigma_m<\varepsilon$（一般情况下可令 $\varepsilon=0.01\sim0.05$），此时有：

$$b_{mi} = A_in_i = \sqrt[k]{\frac{Q_i}{Q_1}} \cdot n_i \approx m$$

即有

$$\frac{Q_i}{Q_1} \approx \left(\frac{m}{n_i}\right)^k$$

由于此时 A_in_i 最接近正整数 m，所以在给定的误差范围内已符合巴内托定律，因此可以把矩阵 B 的第 m 行作为含油气油田规模序列的预测模型。

⑤ 把预测模型序列 $b_{mi}(i=1,2,\cdots,t)$ 中的每个元素除以 A_i，则可得到含油气区中已发现油田储量 $Q_1,Q_2\cdots,Q_t$ 在预测的油田规模序列中的序号 n_i，即：

$$n_i = \frac{b_{mi}}{A_i} \qquad (i = 1,2,\cdots,t) \tag{15-15}$$

式中 n_i——已发现的第 i 个油田的储量在预测的油田规模序列中的序号；

b_{mi}——矩阵 B 中的第 m 行的第 i 列元素；

m——矩阵 B 的行号 m，是已发现的最大油田储量 Q_1 在预测的油田规模序列中的序号，$m\geqslant1$。

⑥ 含油气区中已发现的任何一个油田储量 $Q_i(i=1,2,\cdots,t)$，乘以预测序号 n_i 的 k 次方幂，则为预测的最大（第一号）油田储量 $\hat{Q}_{imax}$。如果所有已发现油田的储量都是可靠的，则应

以所有已发现油田的储量推算 $\hat{Q}_{imax}$ 的平均值，作为含油气区中预测的最大油田储量，即：

$$\hat{Q}_{imax} = \frac{1}{t}\sum_{i=1}^{t} Q_i n_i^k \tag{15-16}$$

⑦ 用预测的最大油田储量 $\hat{Q}_{max}$ 除以 $1^k, 2^k, \cdots$，则得到探区中预测的油田规模序列 $\hat{Q}_j$，即：

$$\hat{Q}_j = \frac{\hat{Q}_{max}}{j^k} \qquad (j = 1, 2, \cdots, p) \tag{15-17}$$

当预测的油田规模序列中第 $p+1$ 个储量 $\hat{Q}_{p+1} < Q_{min}$ 时，可以截断预测序列。Q_{min} 为人为规定的在当时经济技术水平下最小经济油田的储量值。

⑧ 预测全探区总的石油储量(或资源量)$S\hat{Q}$：

$$S\hat{Q} = \sum_{j=1}^{p} \hat{Q}_j = \sum_{j=1}^{p}\left(\frac{\hat{Q}_{max}}{j^k}\right) \tag{15-18}$$

⑨ 按 $k = \text{tg}\ 115° \sim \text{tg}\ 155°$ 范围内的步长，分别计算 s 个预测的油田规模序列 $\hat{Q}_r$ 之中与已发现油田对应的预测值 $\hat{Q}_{ri}$，再计算每个序列中已发现油气田的实际储量 $\hat{Q}_i$ 与所预测的储量之间的标准差 σ_r：

$$\sigma_r = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(Q_i - \hat{Q}_{ri})^2} \qquad (r = 1, 2, \cdots, s) \tag{15-19}$$

式中 Q_i——含油气区中已发现的第 i 个油田的实际储量；

$\hat{Q}_{ri}$——第 r 个预测序列中，与已发现的第 i 个油田对应的预测值。

最后在 s 个预测序列中，选定 σ_r 的值为最小的序列作为预测的油田规模序列。

上述的计算结果是经过数学运算后得出的预测值，是否符合实际的地质情况，还需要由熟悉含油气地区地质情况的地质学家们商榷。

4. 算例

某探区在地质构造上属于一个独立的地质凹陷，面积较小，经过勘探发现 4 个小油田，石油地质储量分别是 149.143、61.567、34.375、27.277(10^4t)。

由于该凹陷是个新探区，所以难以确定油田规模序列的系数 k，故需要通过多次拟合计算才能确定 k 值。为了叙述方便，这里把第⑧步的计算结果 $k = -\text{tg}\ 120° = 1.7321$ 在此引用以作示范。

① 把已发现的 4 个油田，按储量由大到小排列如下：

$$Q_1 = 149.143(10^4\text{t}) \qquad Q_2 = 61.567(10^4\text{t})$$

$$Q_3 = 34.375(10^4\text{t}) \qquad Q_4 = 27.277(10^4\text{t})$$

以其中最大的油田储量 149.143(10^4t)作为推算点。

② 用推算点 Q_1 去除 Q_1, Q_2, Q_3, Q_4，并求所得之商的 k 次方根，得到序列 A_i：

$$A_1 = \sqrt[k]{\frac{Q_1}{Q_1}} = 1.0 \qquad A_2 = \sqrt[k]{\frac{Q_2}{Q_1}} = 0.6$$

$$A_3 = \sqrt[k]{\frac{Q_3}{Q_1}} = 0.428\,6 \qquad A_4 = \sqrt[k]{\frac{Q_4}{Q_1}} = 0.375$$

③ 将 $A_i(i=1,2,3,4)$乘以某一正整数，使其乘积值 A_in_i 最大限度地接近下面矩阵的行号，并记入下面矩阵 B：

$$B=\begin{bmatrix}1.0 & 1.2 & 0.8572 & 1.125\\ 2.0 & 1.8 & 2.143 & 1.875\\ 3.0 & 3.0 & 3.000 & 3.000\end{bmatrix}\begin{matrix}b_{1i}\approx 1\\ b_{2i}\approx 2\\ b_{3i}\approx 3\end{matrix}$$

矩阵 B 中的第 1 行各元素是由 A_i 与 n_i 相乘得到的，其中 n_i 分别为：$n_1=1$，$n_2=2$，$n_3=2$，$n_4=3$。即：

$$b_{11}=1.0\times 1=1.0 \qquad b_{12}=0.6\times 2=1.2$$

$$b_{13}=0.4286\times 2=0.8572 \qquad b_{14}=0.375\times 3=1.125$$

矩阵中第 2 行各元素为：

$$b_{21}=1.0\times 2=2.0 \qquad b_{22}=0.6\times 3=1.8$$

$$b_{23}=0.4286\times 5=2.143 \qquad b_{24}=0.375\times 5=1.875$$

矩阵中第 3 行各元素为：

$$b_{31}=1.0\times 3=3.0 \qquad b_{32}=0.6\times 5=3.0$$

$$b_{33}=0.4286\times 7=3.000 \qquad b_{34}=0.375\times 8=3.000$$

矩阵 B 计算到第 3 行时，标准差 $\sigma_3=0.00063$，即可以认为已符合巴内托定律，因面可以把第 3 行作为油田序列规模的预测模型。

④ 预测模型序列 $b_{3i}(i=1,2,3,4)$中的每个元素除以相应的 $A_i(i=1,2,3,4)$，则得到已发现的 4 个油田储量 Q_1、Q_2，Q_3，Q_4 在预测的油田规模序列中的序号：

$$n_1=\frac{3.0}{1.0}=3 \qquad n_2=\frac{3.0}{0.6}=5 \qquad n_3=\frac{3.0}{0.4286}=7 \qquad n_4=\frac{3.0}{0.375}=8$$

这说明已发现的 4 个油田储量 Q_1，Q_2，Q_3，Q_4 在预测的油田规模序列中，序号分别为 3，5，7，8 号。

⑤ 4 个已发现油田的储量 Q_1，Q_2，Q_3，Q_4 分别乘以预测序号的 $k(k=1.7321)$次方幂，即 3^k，5^k，7^k，8^k，则得到 4 个预测的最大油田储量：

$$\hat{Q}_{1\max}=149.143\times 3^k=1000.0026$$

$$\hat{Q}_{2\max}=61.567\times 5^k=999.999$$

$$\hat{Q}_{3\max}=34.375\times 7^k=999.999$$

$$\hat{Q}_{4\max}=27.277\times 8^k=999.987$$

用上述 4 个预测值的平均值 1000(10^4t)，作为含油气区中预测出的最大油田储量 $\hat{Q}_{\max}$。

⑥ $\hat{Q}_{\max}$ 分别除以 1^k，2^k，…，则得到含油气区中预测的油田规模序列 $\hat{Q}_1$，$\hat{Q}_2$，…，$\hat{Q}_p$。这里暂定最小经济油田的储量 $\hat{Q}_{\min}=10(10^4\text{t})$，则得到如下预测结果($10^4$t)：

$$\hat{Q}_1=1000.000 \qquad \hat{Q}_2=301.022 \qquad \hat{Q}_3=149.142(Q_1)$$

$$\hat{Q}_4=90.615 \qquad \hat{Q}_5=61.567(Q_2) \qquad \hat{Q}_6=44.895$$

$$\hat{Q}_7=34.375(Q_3) \qquad \hat{Q}_8=27.277(Q_4) \qquad \hat{Q}_9=22.243$$

$$\hat{Q}_{10}=18.533 \qquad \hat{Q}_{11}=15.713 \qquad \hat{Q}_{12}=13.513$$

$$\hat{Q}_{13} = 11.765 \qquad \hat{Q}_{14} = 10.348$$

用上述预测结果在双对数纸上展点作图，点的连线成一条直线，其斜率为－1.7321，见图15-3。图中的黑点为已知油田，空圈为预测的油田。

⑦ 含油气区的石油储量总和$S\hat{Q}$为：

$$S\hat{Q} = \sum_{j=1}^{14} \hat{Q}_j = 1801.004(10^4 t)$$

⑧ 为预测该含油气区的油田规模序列，总共作了9次拟合计算。当$\theta = 120°$，即$-\text{tg}\,120° = 1.7321$时，拟合效果最佳，已发现的4个油田储量与所预测的储量之间的标准差很小，$\sigma_r = 0.00063$，所以被选定为预测序列。

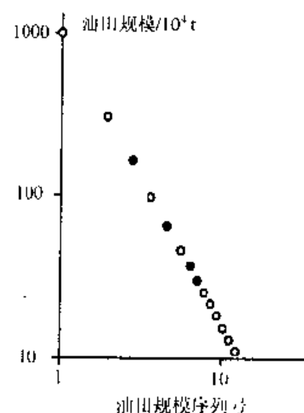


图15-3 某含油气地区的油田规模序列

上述计算结果，经过熟悉含油气区地质情况的地质家们讨论，认为基本符合实际情况。

三、干酪根降解法

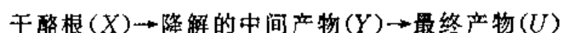
在本书第十二章中，曾结合盆地模拟论述了化学动力学方法，这里所介绍的干酪根降解法是和盆地模拟过程无关的计算生烃量的一种方法。有机物质随沉积物埋藏在地下，在适当的条件下形成化学结构很复杂的干酪根，它是带有许多官能团、分子量很大的具有三度空间结构的有机物质。其后，在适当的温度、压力、催化物质作用下，干酪根将逐步降解生成油气，这就是干酪根降解生成油气的基本原理，也是目前有机生油学说的主要论点。

蒂索(Tissot. B. P)等人于1969年首次把化学动力学理论应用到生油理论研究中，提出了干酪根降解生油的数学模型，把生油理论研究工作向前推进了一步。

1. 干酪根降解生油的基本概念

(1) 干酪根降解生油的两个阶段

蒂索将干酪根在温度和时间因素作用下，向油气转化的过程分为两个阶段，即：



干酪根(X)→降解的中间产物(Y)→最终产物(U)

这一演化过程的中间产物是液态烃(石油)，最终产物是气态烃(天然气)。因此，整个干酪根降解过程可划分为生油及成气两个阶段。

(2) 生油潜量与活化能分布

生油潜量是指干酪根降解成为油气的最大潜力，不同类型的干酪根具有不同的生油潜量。由于干酪根是结构十分复杂的、分子量很大的有机物质，因而其化学活化能不能用单一数值表示。另外由于干酪根中包含多种官能团以及其他杂原子，所以在分子结构中有多种类型的键合。众所周知，各种键合发生反应时的活化能是各不相同的，即使是同一种键合，由于相邻官能团不同，活化能也不相同。因此，研究干酪根的化学反应性能，不能只用单一活化能数值，而要用由多种不同的活化能数值构成一个活化能密度分布来表示。

干酪根的活化能分布，实际上是由各种类型键合活化能数值描述的离散型密度分布。然而，要测定每个单一类型键合活化能是有困难的，因此，常用具有不同活化能反应物质的数量来表示。

根据蒂索等人在实验室对实际样品的测定结果，干酪根所含键合的活化能分布范围在几

kcal/mol 到 80kcal/mol 之间。他们研究了Ⅰ、Ⅱ、Ⅲ型干酪根的生油潜量及活化能分布，以活化能的 6 个离散值代替密度分布，见表 15-2 和图 15-4。

表 15-2　三种类型干酪根的活化能分布及生油潜量

活化能		干酪根类型					
类型 E_{1i}	平均值 /(kcal/mol)	Ⅰ型		Ⅱ型		Ⅲ型	
		X_{i0}	A_{1i}	X_{i0}	A_{1i}	X_{i0}	A_{1i}
E_{11}	10	0.024	4.75×10^{4}	0.022	1.27×10^{5}	0.023	5.2×10^{3}
E_{12}	30	0.064	3.04×10^{16}	0.034	7.47×10^{16}	0.053	4.20×10^{16}
E_{13}	50	0.136	2.28×10^{25}	0.251	1.48×10^{27}	0.072	4.33×10^{25}
E_{14}	60	0.152	3.98×10^{30}	0.152	5.52×10^{29}	0.091	1.97×10^{32}
E_{15}	70	0.347	4.47×10^{31}	0.116	2.04×10^{35}	0.049	1.20×10^{33}
E_{16}	80	0.172	1.10×10^{34}	0.120	3.80×10^{35}	0.027	7.56×10^{31}
$X_0=\sum_i X_{i0}$		0.895		0.695		0.313	
Y_0		0.051		0.035		0.018	

E_{1i}——成油阶段干酪根中第 $i(i=1,2,\cdots,6)$ 种物质的活化能；

X_0——干酪根的生油潜量，是指单位重量干酪根能生成油气的最大比值，其中Ⅰ、Ⅱ、Ⅲ型干酪根的 X_0 值分别为 0.895、0.695、0.313；

Y_0——积分初值，是指干酪根降解前已转化为烃类的比值，其中Ⅰ、Ⅱ、Ⅲ型干酪根的 Y_0 值分别为 0.051、0.035、0.018。

上述情况说明，不同类型的干酪根具有不同的生油潜量，Ⅰ型最高，Ⅱ型居中，Ⅲ型最低；同一类型的干酪根中，活化能的频率分布也不一样，Ⅰ型干酪根以活化能 70kcal/mol 为主，在单位重量的干酪根中占 0.34。Ⅱ型干酪根以活化能 50kcal/mol 为主，占 0.25。Ⅲ型干酪根以活化能 60kcal/mol 为主，占 0.091。

根据蒂索等人的上述数据，经过模拟计算可知，随着地层温度的增加，干酪根中活化能不同的 6 种物质是按活化能的增大顺序依次发生反应。在图 15-5 中描绘了Ⅱ型干酪根 6 种活化能物质随埋藏深度增加的变化情况，即描述了随着地温增加，干酪根中 6 种物质发生降解的情况，这说明引入活化能密度分布概念的重要性。

关于干酪根活化能的密度分布，从空间上来看，它表示了活化能不同的 6 种物质在干酪根中的相对比例关系。从时间上来看，它体现了活化能不同的 6 种物质在降解生油过程中的依次演变过程。就是说，随着埋藏深度的增加，地温不断升高，在这一过程中，干酪根中活化能最低的物质先发生反应，然后是活化能高的物质发生反应。亦即由于活化能不同，发生降解反应的起始时间有先后之分。当几种物质都发生反应后，就同时存在几种平行反应，而各种物质反应的完成时间也是有先后之分。

(3) 干酪根降解生油的数学模型

假定干酪根降解过程服从阿雷尼乌斯方程，则可建立如下微分方程组：

$$\begin{cases}\dfrac{dX_i}{dt}=-K_{1i}X_i\\ \dfrac{dU_j}{dt}=K_{2j}Y\\ Y=\sum Y_i\\ \sum X_{i0}+\sum Y_{i0}+\sum U_{j0}=\sum X_i+\sum Y_i+\sum U_j\end{cases}\tag{15-20}$$

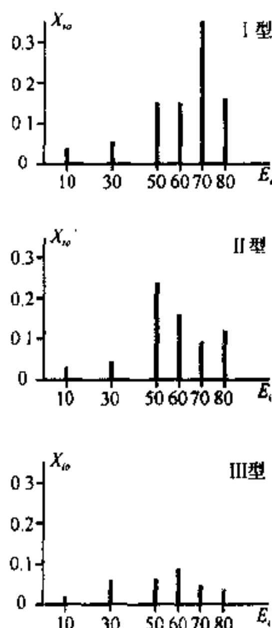


图 15-4 干酪根活化能的密度分布图

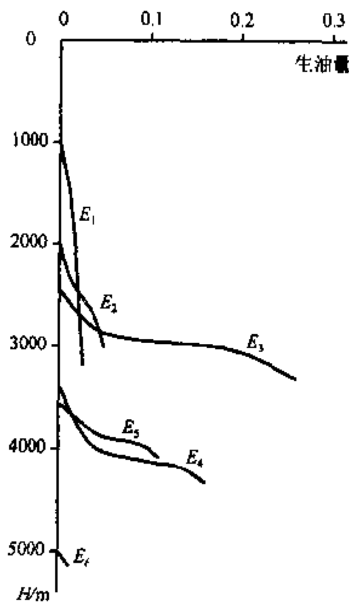


图 15-5 Ⅱ型干酪根降解过程中 6 种活化能物质的降解曲线

$$\begin{cases} K_{1i} = A_{1i}\exp(-\dfrac{E_{1i}}{RT}) \\ K_{2j} = A_{2j}\exp(-\dfrac{E_{2j}}{RT}) \end{cases} \tag{15-21}$$

式(15-20)及式(15-21)中：

t——时间，Ma；

X_i——干酪根中第 i 种活化物质在时刻 t 的数量比例；

K——反应速率，Ma^{-1}；

A——频率因子，Ma^{-1}；

E——活化能，kcal/mol；

K_{1i}——成油阶段干酪根中第 i 种活化能物质的反应速率；

A_{1i}——成油阶段干酪根中第 i 种活化能物质的频率因子；

E_{1i}——成油阶段干酪根中第 i 种物质的活化能；

K_{2j}——成气阶段的反应速率，当最终降解产物为一种气体时 $j=1$；

A_{2j}——成气阶段的频率因子；

E_{2j}——成气阶段的活化能；

R——气体常数，$R=1.986$ cal/mol；

T——绝对温度，K；

Y——生油量，mkg 碳/mkg 有机碳；

U_j——生气量，mkg 碳/mkg 有机碳；

X_{i0}——时间 $t=0$ 时，干酪根中第 i 种物质的初值；

Y_{i0}——时间 $t=0$ 时，液态烃物质的数量；

U_{j0}——时间 $t=0$ 时，气态烃物质的数量。

绝对温度 T 由式(15-22)计算：

$$T = G \cdot v \cdot t + T_0 + 273 \quad (15\text{-}22)$$

式中 G——地温梯度，℃/hm；

v——沉降速度，m/Ma；

t——时间，Ma；

T_0——地表年平均温度，℃。

上述数学模型是对生油机理的简要描述，尽管还不十分完善，如未考虑地层压力、催化条件等，但它毕竟是从机理上导出的理论模型。这一模型把时间、地温、生油量三者间的关系定量地联系起来，可以定量计算出油气生成数量。

需要指出，方程组(15-20)中前两式的等号右侧的反应速率 K 是随时间改变的，所以不能用简单的积分方法求解，而应采用数值积分方法求解。

2. 油气生成量的计算步骤

(1) 确定计算参数

① 地质参数。包括生油岩地质时代及最大沉降深度/m、沉降速度/m/Ma、地温梯度/℃/hm、地表年平均温度/℃、生油岩分布面积/km^2、生油岩厚度/m、生油岩密度(0.023×10^8t/$km^2\cdot$m)。

② 地球化学及热动力学参数。包括干酪根含量(可用有机碳含量表示)、干酪根类型(计算时选用相应的活化能 E 及频率因子 A)、干酪根生油潜量 X_0 及积分初值 Y_0。

(2) 求解数值积分

求解式(15-20)及式(15-21)，可采用龙格—库塔法，求解第 $n+1$ 步的积分公式如下：

$$Y_{n+1} = Y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$k_1 = Hf(X_n, Y_n)$$

$$k_2 = Hf\left(X_n + \frac{H}{2}, Y_n + \frac{k_1}{2}\right)$$

$$k_3 = Hf\left(X_n + \frac{H}{2}, Y_n + \frac{k_2}{2}\right)$$

$$k_4 = Hf(X_n + H, Y_n + k_3)$$

上述公式中的 H 为时间步长(Ma)，$f(X,Y)$为微分方程的函数式。

(3) 计算过程

由于干酪根降解是由生油及气两个阶段组成的，所以实际计算过程也分两步。从什么时间开始进行生气量计算是首先需要解决的问题。

确定生气阶段的起始时刻是一个比较复杂的问题，因此，应当综合考虑勘探地区的各种实际资料，例如镜煤反射率、气层在探区中实际出现的深度等资料来确定。实际计算时，应从干酪根降解过程中因升温而增加反应速度的基本原理来判断。即根据温度升高，反应速度增大的原则，如果温度继续增高时，生油速度反而减慢了，便认为增加的热能消耗于液态烃的裂解过程

中。因此，可将生油阶段的活化能 E_2，频率因子 A_2，进行生气量计算。

① 单位生油量计算。以上所说的计算都是指干酪根的降解率(或称转化率)，所以称作单位生油量。

令 M 时刻的生油量为 YB，则有：

$$YB(M) = X_0 + Y_0 - XI \tag{15-23}$$

式中 X_0——生油潜量，$X_0 = \sum X_{i0}$；

Y_0——原始液态烃数量，$Y_0 = \sum Y_{i0}$；

XI——为 M 时刻剩余的干酪数量，由数值积分方法得到，$XI = \sum X_i$ 。

② 单位生气量计算。在干酪根降解初期直至降解生油速度达到最大值之前，可以忽略不计非降解的其他成因生成的气体，例如生物成因的气体。这里的生气量是指从生气点开始后的降解生气量，即计算由液态烃(石油)裂解生成的气体数量。

进入生气阶段后，干酪根降解系统进入了三相状态，即有固体的干酪根、液态的石油及气态的天然气。令 M 时刻的生气量为 $YG(L)$，则有：

$$YG(L) = YB(M) - YU(L) \tag{15-24}$$

其中：$YG(L)$——生气量；

$YB(M)$——M 时刻干酪根降解产物的总量；

$YU(L)$——降解三相系统中的液态烃数量。

③ 降解率计算。在 M 时刻，降解率 $KR(M)$为：

$$KR(M) = \frac{X_0 - \sum X_i}{X_0} \tag{15-25}$$

④ 生成速度计算。在 M 时刻，生成速度 $VB(M)$为：

$$VB(M) = \frac{YB(M) - YB(M-1)}{HW} \tag{15-26}$$

式中 HW——时间增量，即输出步长。

⑤ 生油岩的生油量计算。数值积分的结果可以给出单位干酪根的降解率，据此便能估算出勘探地区中某一生油岩分布地区的总生油量 Q：

$$Q = S \cdot H \cdot D \cdot C \cdot X_0 \cdot KR \tag{15-27}$$

式中 S——生油岩分布面积，km^2；

H——生油岩厚度，m；

D——生油岩密度，一般取 $D=0.023 \times 10^8 t/km^2 \cdot m$；

X_0——干酪根生油潜量(包括生油潜量和原始液态烃)；

KR——降解率，%。

Q——生油量，$10^8 t$。

3. 算例

赵旭东等人按上述方法估算我国某小型陆相盆地的总生油量。计算时借用了蒂索发表的热动力学参数，采用自动变步长方法进行积分计算。

(1) 地质参数

① 生油岩时代为第三纪；

② 最大沉降深度 H=5000 m；

③ 沉降速度 V=100 m/Ma；

④ 地温梯度 $G=3.6℃/hm$；

⑤ 地表年平均温度 $T_0=14℃$；

⑥ 生油岩分布面积 $S=347.4\ km^2$；

⑦ 生油岩密度 $D=0.023\times10^8 t/km^2\cdot m$；

⑧ 生油岩厚度 $H=409.5\ m$；

⑨ 有机碳量 $C=1.2\%$。

(2) 地球化学及动力学参数

① 干酪根类型为Ⅰ型；

② 生油阶段的 X_{i0}、A_{1i}、X_0、Y_0 选用了表 15-2 中Ⅰ型干酪根的数据；

③ 生气阶段的活化能 E_2，经过多次试算，认为选用 $E_2=70$ kcal/mol，$A_2=0.2\times10^{36}$为宜。

(3) 温度计算

在阿雷尼乌斯公式中，需要使用 T(绝对温度)，从有机质演化角度看，温度 T 是时间 t 的函数，即：

$$T=G\cdot v\cdot t+T_0+273$$
$$R\cdot T=R\cdot G\cdot v\cdot t+R(T_0+273)$$

对该地区来说：

$$R\cdot G\cdot v=7.150$$
$$R(T_0+273)=569.982$$

(4) 计算结果

按上述参数计算，干酪根降解生油演化及降解率见图 15-6 及图 15-7。从图 15-6 及图 15-7 可以看出，该地区干酪根降解生烃过程可分为三个阶段：

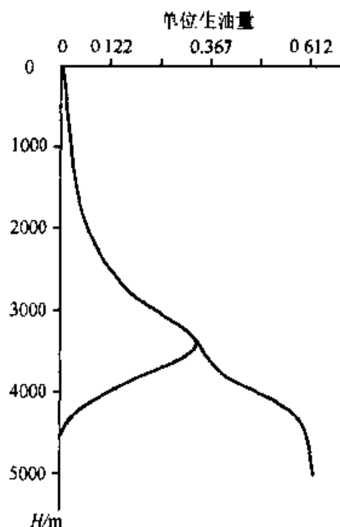


图 15-6 干酪根降解生烃演化图

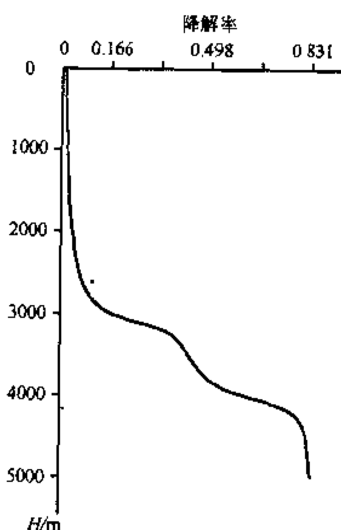


图 15-7 干酪根降解率图

① 初期生油阶段。从深度 1500m 开始至 2200 m，相当于地温由 54°升至 79.2℃，生油速度由 0.09%增加到 0.37%。

② 主要生油阶段。深度为 2200 m 至 3200m，地温增至 115.2℃，生油速度达到 5.2%。

③ 生气阶段。深度从 3200 m 开始进入生气阶段，干酪根降解系统进入三相状态，液态烃(石油)逐渐减少。至 4400 m 时，地温为 158.4℃，干酪根降解产物全部为气，降解率达到 82%。在深度 4400 m 至 5500 m 时，干酪根在高温条件下直接成气，降解率已达到 83%。

经过计算该地区的生油岩总生油量 $Q=23.75\times10^{8}$t。

最后需要指出，这个算例中借用蒂索发表的动力学参数未必合理。

四、特尔菲法

特尔菲(Delphi)法目前的含义是对同一问题的多种见解或多种判断所进行综合处理的一种方法。在石油资源评价中，特尔菲法是一种客观地综合石油地质专家们的知识、经验或见解的技术。从集思广益的角度看，特尔菲法无疑是有价值的，特别是在探区的早期资源评价时更为有用。

用特尔菲法进行石油资源评价的前提条件是：已有的地质资料是够用的，参加石油资源评价的地质专家是合格的。在这两个基本条件下，得出的评价结论才能被认为是可信的。

1. 特尔菲法的要点

① 要有一名有经验的石油地质专家主持石油资源评价工作，这位专家可称为特尔菲班长。由班长聘请若干名具有丰富经验及渊博知识的石油地质专家，组成勘探地区的资源评价小组。因此，也有人把特尔菲法称作专家评价法。

② 资源评价小组中每个专家使用的评价方法，要由专家本人确定，特尔菲班长不能作任何干预，以期保证每个专家充分发挥自己的经验和才干。

③ 资源评价小组成员之间匿名，以防止因评价组内因存在某些技术权威人士而使评价意见产生倾向性。

④ 评价任务可以一次完成，也可以多次反复进行，即以匿名方式把前一轮每个专家的评价结果"反馈"给所有专家，让每个专家再次认真研究核对，并且重新给出下一轮的评价意见。

⑤ 每个专家的评价意见，最好以各种概率下资源量估计值的形式给出。而特尔菲班长则要根据各位专家的评价意见，综合出最终的评价结论，并以分布函数的形式表示，即应该给出不同置信水平下的石油资源估计值。

2. 特尔菲法的实施步骤

① 确定勘探地区的评价范围，即各位专家进行资源评价的地域范围必须是同一的、唯一确定的。

② 选择有经验的石油地质专家作为特尔菲班长。如果班长认为需要的话，也可选定若干名助手协助工作。班长负责主持整个石油资源评价工作，助手只对班长负责，作好班长委托的具体工作。

③ 由特尔菲班长负责选聘若干名石油地质专家组成一个石油资源评价小组。这些专家应当有丰富的实践经验、对评价地区的地质情况应当有较为详细的了解。特别是要求这些专家能够坚持实事求是的原则。

所聘请专家的数量，原则上越多越好，但特别要注意专家的质量。对于每位专家，班长可根据他的水平，赋予不同的权，用以确定各位专家在评价组中的作用。当然，这个权值只能由班长自己掌握，而不宜公开。

④ 由班长拟定向各位专家征询评价意见的表格以及征询内容。目前最常用的是一种以概率分布形式提问的征询表。例如，可向专家们提出如下一些问题：

1° 在评价区中发现 10Mt 以上石油资源量的可能性有多大？

2° 发现 50 Mt 以上石油资源量的可能性有多大？

3° 发现 100 Mt 以上石油资源量的可能性有多大？

4° 发现 500 Mt 以上石油资源量的可能性有多大？

或者把问题反过来提出：

1° 在评价区至少能拿到多少石油资源量(即概率为 100%时的资源量)？

2° 在概率为 75%时的石油资源量有多少？

3° 在概率为 50%时的石油资源量有多少？

4° 在概率为 25%时的石油资源量有多少？

5° 最多可以拿到多少石油资源量(即概率趋向 0%时的资源量)？

⑤ 向不熟悉或不习惯用概率方法估算石油资源量的专家们解释或说明用概率估值的意义和方法。一般情况下，多数专家乐于接受以概率形式的提问。个别专家因为不熟悉概率方法，坚持采用单一估值给出评价意见时，特尔菲班长也不必强求非用概率方法给出不行。因为只要经过简单的数学方法处理，单一的点估值也可以满足尔后的计算要求。但应当尽量使专家们不要采用单点估值方法。

⑥ 每个专家在详细了解评价区的地质情况后，根据自己的经验、知识、习惯，确定评价方法，并按特尔菲班长的征询内容回答问题。

⑦ 特尔菲班长将所有专家的评价结果，用约定的算法进行综合，得出对评价区的综合评价结论。

⑧ 由特尔菲班长与各位专家进行单线联系，讨论、商榷各位匿名专家的评价意见以及班长得出的综合评价结论。其目的是尽可能消除重大分歧，以求得尽可能的统一认识。

⑨ 各位专家根据特尔菲班长给出的第一轮评价结论，以及与班长的商榷结果，重新提出修改后的评价意见(或者坚持自己的原有评价意见)。然后回到第 7 步，再由班长综合出第二轮评价结论。如此反复，直到两轮评价结果无明显差别时，特尔菲班长即可认为完成了评价工作。

⑩ 将最终的评价结果，向上级领导或委托单位呈报评价区各种概率下石油资源量的估计值。

3. *特尔菲班长的综合处理方法*

虽然不少书籍、文献中都论述过特尔菲法的要点和步骤，但是，尚未见到有关特尔菲班长如何综合处理各位专家评价意见的具体算法。本书列举的两种综合算法——概率加权法和加权抽样法是按特尔菲法的基本要点及工作步骤拟定的(见《石油数学地质概论》，赵旭东编著)。

下面以一个算例来具体说明在石油资源评价中如何应用特尔菲法。

某沉积盆地经地震及少量钻井证实，该盆地为一个含油气远景区。为估算该盆地的石油资源量，评价单位选聘了特尔菲班长，由班长聘请了 11 名具有丰富经验的石油地质学家组成一个评价小组。特尔菲班长将自己掌握的全部地质资料印发给每个专家。每个专家在消化地质资料的基础上，用不同的找油理论及相应的预测方法，给出了评价意见，见表 15-3。

为了叙述方便，这里约定如下：称第 i 个专家的石油资源量估计值为 Q_i，对应的累积概率为 $AF_i(i=1,2,\cdots,R)$。若第 i 个专家给出了 N 个不同概率下石油资源量的估计值，则第 i 个专家的第 j 点估计值用 Q_{ij}表示($j=1,2,\cdots,N$)。第 i 个专家估计值中的极小值与极大值分别用 $Q_{i\min}$ 和 $Q_{i\max}$ 表示。

前已述及，各位专家应尽量避免用单一的点估计值。但是表 15-3 中的 6～9 号专家，由于不习惯或不熟悉概率估值方法，他们坚持给出单一的点估计值。为了进行尔后的计算，需要进

行某些处理。这里是采用小区间展开进行处理，即认为这个点估计值相当于概率为50%时的估计值。

表 15-3 专家评价意见汇总表

专家号 R	各位专家的权系数 W	资源量估值点数 N	各概率水平下的石油资源量估计值 $Q_{ij}/10^8t$				
			100%	75%	50%	25%	0%
1	1	2	22.22				34.17
2	1	2	31.51				42.06
3	1	2	20.95				30.65
4	1	2	39.63				57.60
5	1	2	28.10				52.24
6	1	1			32.77		
7	1	1			46.00		
8	1	1			70.00		
9	1	1			65.00		
10	1	5	20.15	26.32	27.49	28.90	34.05
11	1	5	30.75	16.12	51.13	54.96	65.17

小区间展开法是把这个点估计值除以某个数C(C>1)，得到一个在数量上远小于估计值Q的Δq，即：

$$\Delta q = \frac{Q}{C} \tag{15-28}$$

再以Q为中心向两侧分别外推Δq，得到$(Q-\Delta q)$及$(Q+\Delta q)$两个值。这两个值可以看作是概率分别为100%及0%区间估计值的端点。

为了区别对待每个专家的作用，特尔菲班长可以用不公开的方式，赋给每个专家以不同的权系数W_i，一般取W_i为正整数。让经验丰富的专家具有较大的权，亦即认为一个有经验的专家所起的作用相当于两或多个专家的作用。如果特尔菲班长对所有专家的评价意见一视同仁，则所有专家的权系数都赋值为1。在我们的实例中所有专家的权系数W_i均等于1。

有了以上准备，特尔菲班长即可以进行综合计算。

(1) 概率加权法

首先在R个专家所给出的所有石油资源量估计值Q_{ij}中，找出最大的估计值Q_{max}及最小的估计值Q_{min}。如果在R个专家中，有些专家的评价值是以点估计值给出的，则要在用小区间展开方法处理后，以$(Q_i-\Delta q_i)$及$(Q_i+\Delta q_i)$代替原来的估计值Q_i，参加Q_{max}及Q_{min}的挑选。

选出的Q_{max}及Q_{min}作为最终评价结论的区间估计值的两个端点，进而再把这一区间分为m个子区间，并求出$m+1$个子区间的分隔值Q_k $(k=1,2,\cdots,m+1)$。

求$m+1$个与石油资源量Q_k相对应概率值的加权平均值AF_k，可按如下公式计算：

$$AF_k = \sum_{i=1}^{R}(AF_i \cdot W_i)/SW \qquad (k=1,2,\cdots,m+1) \tag{15-29}$$

$$AF_i = \begin{cases} 1 & Q_k < Q_{imin} \\ \dfrac{(AF_{ij}-AF_{ij-1})(Q_k-Q_{ij-1})}{(Q_{ij}-Q_{ij-1})} + AF_{ij-1} & Q_{imin} \leqslant Q_k \leqslant Q_{imax} \text{ 且 } Q_{ij-1} \leqslant Q_k \leqslant Q_{ij} \\ 0 & Q_k > Q_{imax} \end{cases}$$

(15-30)

$$(k = 1,2,\cdots,m+1 \quad i = 1,2,\cdots,R \quad j = 1,2,\cdots,N)$$

$$SW = \sum_{i=1}^{R} W_i \tag{15-31}$$

式中 Q_k——第 k 个点的石油资源量估计值；

AF_k——Q_k 的概率值；

R——专家人数；

AF_i——第 i 个专家 Q_k估计值的概率；

W_i——第 i 个专家的权系数；

SW——所有专家的权系数之和；

Q_{ij}——第 i 个专家的第 j 点资源量估计值；

AF_{ij}——Q_{ij}的概率值；

Q_{ij-1}——第 i 个专家的第$(j-1)$点资源量估计值；

AF_{ij-1}——Q_{ij-1}的概率值；

Q_{imax} ——第 i 个专家最大的资源量估计值；

Q_{imin} ——第 i 个专家最小的资源量估计值。

上述计算 AF_i 公式的含义是：当 Q_k 小于 Q_{imin} 时，令第 i 个专家估计值的概率 AF_i 为100%。当 Q_k 大于 Q_{imax} 时，令第 i 个专家估计值的概率 AF_i 为 0%。当 Q_k 大于等于 Q_{imin} 同时小于等于 Q_{imax} 时，用线性插值求出第 i 个专家估计值的概率 AF_i。

最后把计算出来的 $m+1$ 个点(Q_k,AF_k)以分布函数的形式表示出来，这就是根据 R 个专家的评价意见综合出来的最终评价结论。这一结论给出了不同概率下的石油资源量的估计值。

需要指出的是，个别专家的离群估计值，是以小概率出现于最终的评价结论之中，对最终的评价结论影响不大。

(2) 加权抽样法

加权抽样法是利用随机数对 R 个专家给出的石油资源量分布函数进行随机抽样计算。具体的作法是以每个专家的权系数 W_i(W_i 必须是正整数)作为抽样次数，再把 R 个专家的抽样值累加起来，除以 R 个专家的权系数和 SW，得到一个复合抽样值 Q_k，即：

$$Q_k = \sum_{i=1}^{R}\sum_{j=1}^{W_i} Q_{ij}/SW \qquad (k = 1,2,\cdots,g) \tag{15-32}$$

式中 Q_k——第 k 个复合抽样值；

Q_{ij}——第 i 个专家的第 j 次抽样值；

W_i——第 i 个专家的权系数；

SW——R 个专家的权系数之和，$SW = \sum_{i=1}^{R} W_i$。

如此反复进行复合抽样，如果令 $g=2000$，则可得到 2000 个复合抽样值。最后以频率统计法求出石油资源量估计值的分布函数，从而给出各种概率下石油资源量的估计值。

用加权抽样法得到的结果与概率加权法相比，往往会有明显的区别。两种算法的区别在于：概率加权法给出的石油资源量估计值的范围较宽，而加权抽样法给出的范围较窄。所以，加权抽样法具有更强的综合能力，更能代表多数专家的评价意见。

最后还要指出，特尔菲班长的综合方法绝非限于上面讲的两种方法，应当根据实际情况，选定最合适的综合方法。

4. 算例

对于表 15-3 中 11 位专家对沉积盆地所作的评价意见，特尔菲班长首先要对第 6～9 号 4 位专家的点估计值进行小区间展开。如果令式(15-28)中的 $C=20$，则以 C 除以各专家的点估计值得到 Δq 值，例如其中 8 号专家的点估计值是 70×10^8t，则有：

$$\Delta q=\frac{70}{20}=3.5(10^8\text{t})$$

所以，第 8 号专家估计值展开后的端点为：

$$Q_{8\min}=70-3.5=66.5(10^8\text{t})$$

$$Q_{8\max}=70+3.5=73.5(10^8\text{t})$$

同样，对第 6、7、9 号专家点估计值也要分别展开为区间估计值。

(1) 按概率加权法计算

按概率加权法计算时，$SW=11$。而 11 位专家估计值中，最大及最小的估计值为：

$$Q_{\max}=73.5(10^8\text{t})$$

$$Q_{\min}=20.15(10^8\text{t})$$

按式(15-29)、(15-30)、(15-31)计算，该沉积盆地不同概率下石油资源量的估计值见表 15-4，其分布函数见图 15-8。

表 15-4 概率加权法综合的石油资源量估计值表

概率/%	石油资源量/10^8t	概率/%	石油资源量/10^8t	概率/%	石油资源量/10^8t
100	20.1500	65	33.0812	30	48.4494
95	23.6947	60	34.0136	25	52.0911
90	26.0960	55	36.7326	20	57.4173
85	27.5879	50	40.1502	15	64.1822
80	29.0688	45	43.4057	10	67.0704
75	31.0514	40	45.2621	5	69.6500
70	32.1487	35	46.8505	0	73.5000

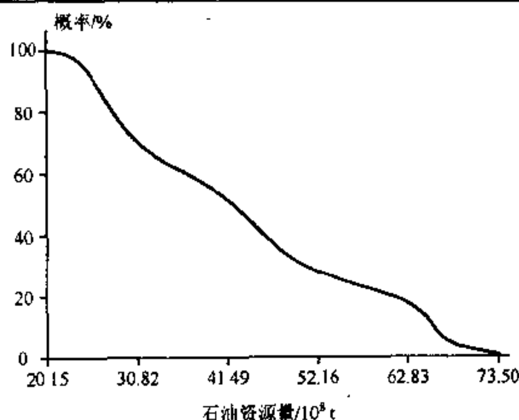


图 15-8 概率加权法综合的石油资源量分布函数

(2) 按加权抽样法计算

用加权抽样法计算的过程见图 15-9。

加权抽样法的计算结果见表 15-5，其分布函数见图 15-10。

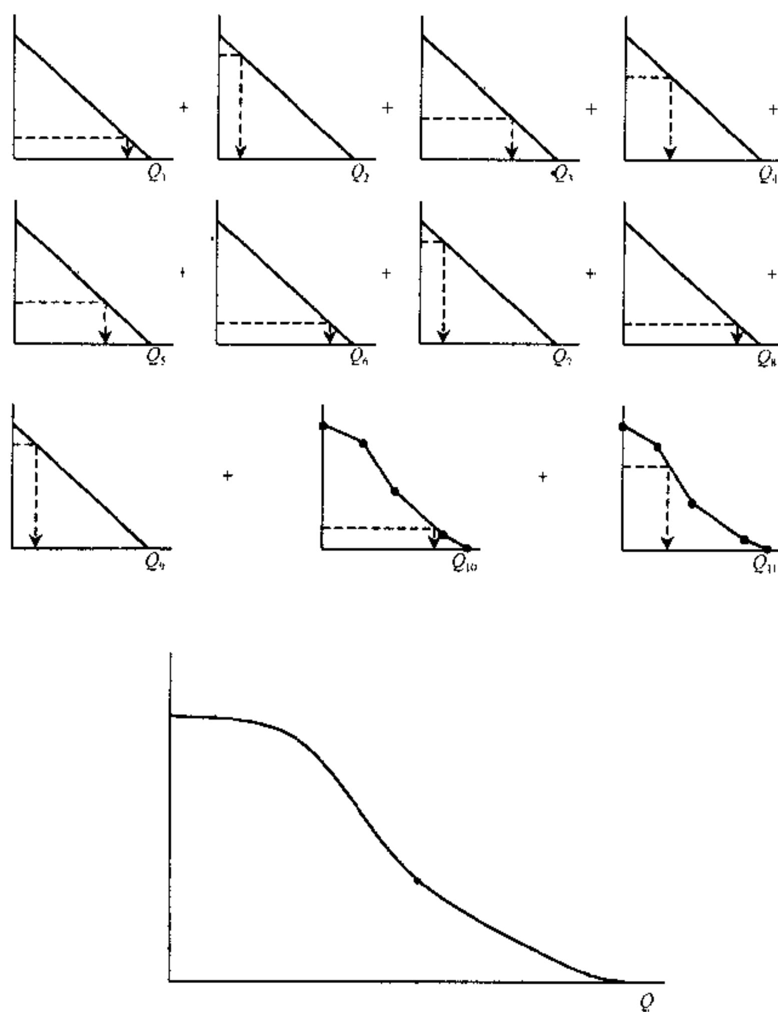


图 15-9　加权抽样法计算过程的示意图

表 15-5　加权抽样法综合的石油资源量估计值表

概率 /%	石油资源量 /10^8t	概率 /%	石油资源量 /10^8t	概率 /%	石油资源量 /10^8t
100	38.6535	65	42.3428	30	43.5027
95	40.6200	60	42.5183	25	43.7017
90	41.1269	55	42.6797	20	43.9087
85	41.4547	50	42.8449	15	44.1566
80	41.7402	45	43.0084	10	44.4679
75	41.9646	40	43.1646	5	44.9182
70	41.1699	35	43.3341	0	47.0325

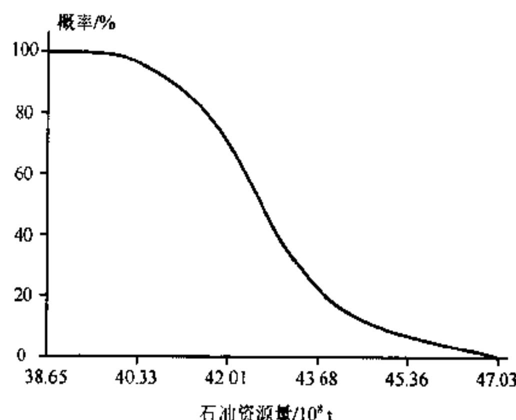


图 15-10　加权抽样法综合的石油资源量分布函数

§2　含油气有利地带的预测方法

目前众多的石油资源评价方法中，绝大多数的方法都属于石油资源量的预测方法，而预测探区含油气有利地带的方法很少。这也是现阶段石油资源评价工作中的一个薄弱环节。研究含油气有利地带的预测方法，不仅有利于提高评价结论的可靠性，而且可以指导探区当前的勘探工作。本节介绍两种含油气有利地带的预测方法。

一、模糊集合综合评介法

地质学中有许多含义模糊的概念，例如某一地质单元的含油气远景，某个地质圈闭的含油性等，其概念都是不准确的。因为构成这些概念的研究对象是没有确定边界的模糊体系，而模糊体系若用严格的数学方法处理，则可能会导出不真实的结果。

所谓模糊体系是指一些复杂的实际问题，它们不可能得到准确和明确的解答，因而需要用描述和分析的方法，来适应那些不准确的知识接界，或者适应我们主观上对实际问题有关价值的判断或评价。

石油勘探阶段，特别是早期阶段，经过地面地质调查或地球物理勘探后，发现了一批地质圈闭，此时勘探人员最关心的问题就是这批圈闭中哪些是含油的，哪些是可能含油的，哪些是不含油的，这就是通常所说的地质圈闭的含油性评价问题。某个地质圈闭中有没有石油，有多少石油储量，原本是地质历史演化的结果。我们在研究这一地质圈闭有多少石油储量时，虽然该圈闭的石油储量在客观上早已确定，但是，限于目前勘探手段所构成的观测系统的技术水平及观测精度，而使我们所能得到的这批地质圈闭含油性的映象却是一个模糊体系。因此，勘探人员依据这一模糊映象，就不可能准确或明确地回答地质圈闭的含油性问题。按以往的常规研究方法，通常是由观测到的地质信息加上勘探人员的实际经验，对每个地质圈闭进行打分，用以描述和分析地质圈闭与控制油气形成的因素之间关系，以适应人们对地质圈闭含油性评价的主观要求，这就是通常所说的按相对好坏给出地质圈闭含油性的排队评价。

美国控制论专家查德(L. A. Zaden)于 1965 年首先提出“模糊集合”的概念，对模糊体系用数学方法进行描述，创立了一个新的数学分支模糊数学。模糊数学是研究和处理模糊体系规律

性的理论和方法，它把普通集合论只取 0 或 1 两个值的特征函数，推广到 [0,1]区间上取值的隶属函数。把绝对的属于或不属于的“非此即彼”扩展为更加灵活的渐变关系，因而便于把“亦此亦彼”中介过渡的模糊概念用数学方法处理。尽管目前模糊数学还不完善，在数学界也没有得到普遍承认，但是，它的思想方法与地质圈闭含油性的评价思路却十分相近。这就是以模糊数学方法对地质圈闭含油性进行综合评价的出发点。

应用模糊数学方法对地质圈闭含油性进行综合评价时，应考虑到以下四个问题：

① 与地质圈闭含油性有关的多个控制油气形成的地质因素之间，可能存在多层次的结构关系。例如，地质圈闭的含油性通常决定于生油条件、储油条件、盖层条件等。而这些基本地质条件，又由在当时勘探程度下可能取得的若干个次一级地质因素构成、例如生油条件可能与地质圈闭所处的生油条件分区、生油岩厚度、生油岩的地球化学指标等地质因素有关。

② 对地质圈闭含油性进行综合评价时，对各个地质因素所起的作用很难给出确切的估计值，一般凭经验确定，通常可用权重分配表示。

③ 设 $X=\{x\}$ 是给定的论域，若论域 X 中任何一个元素 x 都有一个 $\mu(x)$ 与之对应，并且满足 $0\leqslant\mu(x)\leqslant1$，则称 $\mu(x)$ 为隶属函数。需要指出，对如何建立隶属函数，目前还是模糊数学尚未完全解决的理论问题。在实际应用中，建立令人信服的隶属函数十分困难，常以经验公式或者评语级别代替。

④ 矩阵合成运算中，仅用取大取小算子将会丢失很多信息，而使评价结果过于单调，甚至难于鉴别地质圈闭含油性的优劣。

1. 地质圈闭的综合评价方法

地质圈闭的含油性是由多种地质条件所决定的。例如，某个探区经过地震勘探已经发现一批地质圈闭，钻探前需要进行圈闭排队，选择含油性最好的地质圈闭首先进行钻探。

如果评价地质圈闭含油性时，引用了 n 项地质因素，则可构成因素集合 U：

$$U=\{U_1,U_2,\cdots,U_i\cdots,U_n\}$$

式中 $U_i(i=1,2,\cdots,n)$ 是集合 U 的元素或子集。当 U_i 是子集时，它可由 n_i 个元素或次一级子集组成，即：

$$U_i=\{U_{i1},U_{i2},\cdots,U_{ij}\cdots,U_{in_i}\}$$

如果评价地质圈闭含油性时，预想分为 m 个级别，则可设评价集合 V，即：

$$V=\{V_1,V_2,\cdots,V_m\}$$

考虑到所引用的每项地质因素在评价地质圈闭含油性时所起的作用不同，可设 A 为因素集合 U 的权重分配，即：

$$A=\{A_1,A_2,\cdots,A_i\cdots,A_n\}$$

式中 A_i 是 A 的元素或子集。当 A_i 是子集时它可由 n_i 个元素或次一级子集组成：

$$A_i=\{A_{i1},A_{i2},\cdots,A_{ij}\cdots,A_{in_i}\}$$

这里要求：

$$\sum_{i=1}^{n}A_i=1,\sum_{j=1}^{n_i}A_{ij}=1$$

从 U 到 V 的一个模糊映射 $R(U_i)$ 叫作单项因素评价，$U_i\in U$ 时有：

$$R(U_i)=(r_{i1},r_{i2},\cdots,r_{im})$$

对地质圈闭的含油性进行综合评价时，所引用的地质因素有时有准确的定量数据，有时只有相对关系或者定性描述。为了统一起见，在此一律采用相对评语表示子集 $R(U_i)$。对于定量

数据，可以通过等级变换转化为相对评语。

如评价集合分为好、中等、差3个级别时，可按表15-6中的评语级别表示子集 $R(U_i)$。与此类似，非评价集合分为好、较好、中等、较差、差5个级别时，可按表15-7中的评语级别表示子集 $R(U_i)$。

如果 U_i 是 U 的元素，则可由 n 个模糊映射 $R(U_i)$ 组成综合评价变换矩阵 R：

$$R = [r_{ij}]_{n\times m} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}$$

表15-6　3个级别的评语表

评语＼级别	−1	0	1
好	0	0.2	0.8
中等	0.25	0.5	0.25
差	0.8	0.2	0

表15-7　5个级别的评语表

评语＼级别	−2	−1	0	1	2
好	0	0	0	0.2	0.8
较好	0	0	0.2	0.6	0.2
中等	0	0.25	0.5	0.25	0
较差	0.2	0.6	0.2	0	0
差	0.8	0.2	0	0	0

如果 U_i 是 U 的子集，可由 n_i 个模糊映射 $R(U_{ij}) = (r_{ij1}, r_{ij2}, \cdots, r_{ijm})$ 组成单项地质因素的综合评价变换矩阵 R_i：

$$R_i = [r_{ijk}]_{n_i\times m} = \begin{bmatrix} r_{i11} & r_{i12} & \cdots & r_{i1m} \\ r_{i21} & r_{i22} & \cdots & r_{i2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{in_i1} & r_{in_i2} & \cdots & r_{in_im} \end{bmatrix}$$

若 U_i 是 U 的元素，地质因素的权重分配 A 与综合评价变换矩阵 R_h 的合成 B_h 称为第 h 个地质圈闭的综合评价，即：

$$B_h = A \bigcirc R_h \qquad (h = 1, 2\cdots, p) \tag{15-33}$$

若 U_i 是 U 的子集，而 U_{ij} 是 U_i 的元素，则首先要计算次一级地质因素的综合评价 B_{hi}，即：

$$B_{hi} = A_i \bigcirc R_{hi} \qquad (h = 1, 2, \cdots, p \quad i = 1, 2, \cdots, n) \tag{15-34}$$

式(15-34)中，B_{hi} 的脚码 i 代表地质因素的编号，h 代表地质圈闭的编号，◯为合成算子。次一级地质因素综合评价的计算结果要作为上一级综合评价变换矩阵的一行。

最后可用下面的式(15-35)求得每个地质圈闭含油性的综合评价值 D_h，即：

$$D_h = B_h C' \qquad (h = 1, 2, \cdots, p) \tag{15-35}$$

式(15-35)中的C'是等级矩阵的转置矩阵。

当评价集合为好、中等、差3个级别时，可令$C=[-1 \quad 0 \quad 1]$。当评价集合分为好、较好、中等、较差、差5个级别时，可令$C=[-2 \quad -1 \quad 0 \quad 1 \quad 2]$，其余类推。

求出p个地质圈闭含油性的综合评价D_h后，则可按D值的大小进行地质圈闭含油性的排队，最后得到p个地质圈闭含油性相对好坏的次序。对于一个勘探地区，可以按计算得到的排队次序作为地质圈闭钻探的先后顺序。

2. 矩阵合成时的运算规则

普通集合乘积矩阵的第i行第j列的元素值，等于左侧矩阵第i行元素与右侧矩阵第j列元素对应项乘积的代数和。但是，模糊矩阵合成的算子较多，除包括普通集合的乘法算子外，还可以根据实际需要定义多种其他算子。

(1) 基本算子

假设a、b为模糊集合中的两个元素，这里定义了如下几种基本算子：

①
$$a \vee b = \max(a,b) \tag{15-36}$$
表示从a、b两个元素中选择数值大的元素值作为$a \vee b$的运算结果。

②
$$a \wedge b = \min(a,b) \tag{15-37}$$
表示从a、b两个元素中，选择数值小的元素值作为$a \wedge b$的运算结果。

③
$$a \cdot b = ab \tag{15-38}$$
表示a与b两个元素的乘积值，运算规则与普通乘法运算一致。

④
$$a \oplus b = \min(1, a+b) \tag{15-39}$$
表示从1和$a+b$中，选择数值小的元素作为$a \oplus b$的运算结果。

(2) 矩阵合成时的运算方法

地质因素的权重分配A与综合评价变换矩阵R的合成称为地质圈闭的综合评价B，即：

$$A \bigcirc R = B = (b_1, b_2, \cdots, b_m)$$

在矩阵合成运算时，可按实际需要选用合适的算子搭配方法。常用的算子搭配方法有如下4种：

① 取小取大算法。简记为$(\wedge, \vee)$，合成矩阵B中的元素b_j的计算方法如下：

$$b_j = \bigvee_{i=1}^{n} (a_i \wedge r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{15-40}$$

②乘积取大算法。简记为$(\cdot, \vee)$，合成矩阵B中的元素b_j的计算方法如下：

$$b_j = \bigvee_{i=1}^{n} (a_i \cdot r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{15-41}$$

③ 取小求和算法。简记为$(\wedge, \oplus)$，合成矩阵B中的元素b_j的计算方法如下：

$$b_j = \bigoplus_{i=1}^{n} (a_i \wedge r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{15-42}$$

④ 乘积求和算法。简记为$(\cdot, \oplus)$，合成矩阵B中的元素b_j的计算方法如下：

$$b_j = \bigoplus_{i=1}^{n} (a_i \cdot r_{ij}) \qquad (j = 1,2,\cdots,m) \tag{15-43}$$

例如：$A=(0.6 \quad 0.3 \quad 0.2)$

$$R = \begin{bmatrix} 0.2 & 0.3 & 0.5 \\ 0.4 & 0.2 & 0.4 \\ 0.1 & 0.6 & 0.3 \end{bmatrix}$$

按$(\wedge, \vee)$计算得： $B=(0.3 \quad 0.3 \quad 0.5)$

按(·,∨)计算得：　$B=(0.12\quad 0.18\quad 0.3)$

按(∧,⊕)计算得：　$B=(0.6\quad 0.7\quad 1.0)$

按(·,⊕)计算得：　$B=(0.26\quad 0.36\quad 0.48)$

3. 算例

① 根据地质调查及地震勘探，在某探区内发现了5个地质圈闭，见表15-8。

表15-8　地质圈闭的各地质因素评语及数据表

地质因素＼圈闭编号		1	2	3	4	5
生油条件		较差	中等	较好	较好	中等
储油条件		中等	较好	中等	中等	较差
盖层条件		较好	中等	中等	较好	好
构造条件	面积/km²	10	20	10	15	20
	幅度/m	100	200	100	150	90
	断层情况	2条	1条	无	无	1条

从表15-8中可看出，由于该探区的勘探程度较低，所以生油条件、储油条件、盖层条件都是由定性评语描述的。构造条件是用地震资料定量描述的。

根据勘探人员的认真分析，各地质因素的权重分配如下：

综合评价
- 生油条件　(0.25)
- 储油条件　(0.25)
- 盖层条件　(0.15)
- 构造条件(0.35)
  - 面　积　(0.4)
  - 幅　度　(0.3)　(子集)
  - 断层情况　(0.3)

构造条件中的面积、幅度、断层情况是定量数据，为了能与生油条件、储油条件、盖层条件的定性评语搭配使用，可按表15-9中规定的标准分为5个等级。该表中的标准是由熟悉探区地质情况的勘探人员讨论商定的。需要指出，每个探区必须根据具体的情况制定分级标准。

表15-9　各构造因素的评语分级标准

地质因素＼评语	差	较差	中等	较好	好
构造面积/km²	<5	5～10	10～30	30～50	>50
构造幅度/m	<50	50～100	100～200	200～300	>300
断层情况	>2条	1～2条	1条	无	无

按表15-9规定的标准，表15-8中的定量数据都可以转换成相应的评语，转换后的评语见表15-10。

这里选用的地质因素是层次关系的，生油条件U_1、储油条件U_2、盖层条件U_3、构造条件U_4，总共4项地质因素构成因素集合U：

$$U=(U_1,U_2,U_3,U_4)$$

其中U_1、U_2、U_3是因素集合中的元素，而构造条件U_4是子集。子集U_4是由构造面积U_{41}、

构造幅度U_{42}、断层情况U_{43}这3项次一级的地质因素组成，即：

$$U_4=(U_{41},U_{42},U_{43})$$

表15-10 地质圈闭的各地质因素评语表

地质因素 \ 圈闭编号		1	2	3	4	5
生油条件		较差	中等	较好	较好	中等
储油条件		中等	较好	中等	中等	较差
盖层条件		较好	中等	中等	较好	好
构造条件	面积/km²	较差	中等	较差	中等	中等
	幅度/m	较差	中等	较差	中等	较差
	断层情况	较差	中等	好	好	中等

由于各项地质因素及次一级构造因素评语都是按5个级别划分的，所以可构成5个级别的评语集合V：

$$V=(V_1,V_2,V_3,V_4,V_5)$$

因为地质因素分为两级，所以也有如下两级相应的权重分配：

$$A=(0.25\quad 0.25\quad 0.15\quad 0.35)$$

$$A_4=(0.4\quad 0.3\quad 0.3)$$

为了对这5个地质圈闭进行排队，首先要从次一级的构造因素，即构造面积、构造幅度、断层情况开始计算。

按表15-7中5个级别的评语，形成矩阵R_{14}，R_{14}是第1个圈闭的第4项地质因素（构造条件）的综合评价变换矩阵，按（·，⊕）乘积求和算法计算时，第一个圈闭的第4项地质因素的综合评价B_{14}为：

$$B_{14}=A_4\bigcirc R_{14}=(0.4\quad 0.3\quad 0.3)\bigcirc\begin{bmatrix}0.2 & 0.6 & 0.2 & 0 & 0\\0.2 & 0.6 & 0.2 & 0 & 0\\0.2 & 0.6 & 0.2 & 0 & 0\end{bmatrix}$$

$$=(0.2\quad 0.6\quad 0.2\quad 0\quad 0)$$

按同样的方法计算，可以得到第2、3、4、5号地质圈闭的构造条件综合评价：

$$B_{24}=(0\quad 0.25\quad 0.5\quad 0.25\quad 0)$$

$$B_{34}=(0.14\quad 0.42\quad 0.14\quad 0.06\quad 0.24)$$

$$B_{44}=(0\quad 0.175\quad 0.35\quad 0.235\quad 0.24)$$

$$B_{54}=(0.06\quad 0.355\quad 0.41\quad 0.175\quad 0)$$

继续进行地质圈闭的综合评价计算时，上面算得的R_{h4}要作为综合评价变换矩阵R_h中的第4行。如果仍以（·，⊕）乘积求和算法计算，则第1个地质圈闭的综合评价B_1为：

$$B_1=A\bigcirc R_1=(0.25\quad 0.25\quad 0.15\quad 0.35)\bigcirc\begin{bmatrix}0.2 & 0.6 & 0.2 & 0 & 0\\0 & 0.25 & 0.5 & 0.25 & 0\\0 & 0 & 0.2 & 0.6 & 0.2\\0.2 & 0.6 & 0.2 & 0 & 0\end{bmatrix}$$

$$=(0.12\quad 0.4225\quad 0.275\quad 0.1525\quad 0.03)$$

按同样方法计算，可以得到第2、3、4、5号地质圈闭的综合评价：

$$B_2 = (0 \quad 0.1857 \quad 0.425 \quad 0.3375 \quad 0.07)$$
$$B_3 = (0.049 \quad 0.247 \quad 0.299 \quad 0.271 \quad 0.134)$$
$$B_4 = (0 \quad 0.12375 \quad 0.3275 \quad 0.38475 \quad 0.164)$$
$$B_5 = (0.071 \quad 0.33675 \quad 0.3185 \quad 0.15375 \quad 0.12)$$

最后，按式(15-35)计算每个地质圈闭的综合评价值 $D_h(h=1,2,3,4,5)$，其中 D_1 为：

$$D_1 = (0.12 \quad 0.4225 \quad 0.275 \quad 0.1525 \quad 0.03)\begin{bmatrix}-2\\-1\\0\\1\\2\end{bmatrix} = -0.45$$

同理可以得到第2、3、4、5号地质圈闭的综合评价值：

$$D_2 = 0.29$$
$$D_3 = 0.194$$
$$D_4 = 0.589$$
$$D_5 = -0.085$$

按以上计算结果，5个地质圈闭的排队次序应为：D_4, D_2, D_3, D_5, D_1，其中第4号地质圈闭的含油气地质条件最好，第1号地质圈闭的含油气地质条件最差。

② 二连盆地中马尼特坳陷至1984年共发现119个局部构造，按地震反射层 T_5、T_7、T_{11}（浅、中、深）收集了如下地质参数：

1° 构造圈闭面积，km^2。

2° 构造圈闭幅度，m。

3° 构造类型定量化，背斜赋值为3，半背及断背斜为2，断块及其他为1。

4° 生油相带是构造圈闭所处的生油凹陷位置，数字化后，Ⅰ类生油区内的构造圈闭赋值为3，Ⅰ至Ⅱ类生油区之间的为2，其他区内的为1。

119个局部构造先按 T_5、T_7、T_{11} 三个反射层计算后，再对局部构造进行最后的综合评价。表15-11列出了119个局部构造中的前20名含油气地质条件较好的局部构造排队顺序及综合评价值。

表15-11 二连盆地马尼特坳陷局部构造的综合评价结果表

排队序号	构造名称	综合评价值	排队顺序	构造名称	综合评价值
1	阿尔善	2.040	11	哈 邦	−0.960
2	哈达图	0.560	12	邦 东	−0.960
3	蒙古林	0.160	13	萨音乌苏	−0.960
4	贡 尼	0.000	14	巴 润	−1.120
5	哈 南	−0.200	15	巴 东	−1.120
6	准沟东	−0.760	16	毛 音	−1.120
7	额尔热傍	−0.760	17	吉 北	−1.120
8	哈 北	−0.800	18	萨东1	−1.120
9	准 沟	−0.960	19	萨东2	−1.120
10	沃布多东	−0.960	20	毛音南	−1.320

二、多种信息叠合评价法

多种信息叠合评价法是对已掌握的探区地质资料进行综合处理的一种方法。通过叠合处理可以得到与含油气有利地带关系密切的综合信息，因而有利于制定探区的勘探方案。

含油气有利地带的预测，不仅仅限于钻探井位的选择，而且应包括全国范围内的有利含油气盆地的选定、沉积盆地内有利地质坳陷的挑选、地质坳陷内有利凹陷或凸起的确定、凹陷或凸起上有利地质圈闭的排队以及地质圈闭上最佳钻探井位的圈定等一系列预测工作。

按传统的石油地质研究方法，为确定含油气有利地带，首先要分别研究生油、储油、盖层、运移、聚集、保存等控制油气形成的基本地质条件，最后再通过综合研究工作进行探区内含油气有利地带的预测。

1. 多种信息叠合评价法的要点

一般情况下，一个新的探区总会有或多或少的地质资料。这些不同类型的地质资料（包括地质数据、地质图件、地质观点），都应当看作是从不同侧面向地质人员提供的寻找有利勘探地带的地质信息。

多种信息叠合评价法的基本思路，是把控制油气形成的各种不同的单一地质因素看作是基础地质信息，表示基础地质信息的地质图件称为基础地质信息图件。由若干个基础地质信息图件经叠合生成的图件称为组合地质信息图件。再由若干个组合地质信息图件经二次叠合生成的图件称为综合地质信息图件。

可见，多种信息叠合评价法的要点可以概括为"图加图出新图"。新图中的信息是由基础地质信息经过逐级多次叠合成的复合信息。

2. 多种信息叠合评价法的实施步骤

(1) 地质数据的归类与分级

地质数据的归类与分级是将已收集的地质数据，形成归类合理的、层次分明的数据体系。一般情况下，可分为两个此层次，即基础地质信息和组合地质信息，同类的基础地质信息构成组合地质信息，再由组合地质信息构成综合地质信息。

例如在某个探区已收集到生油岩厚度、生油岩有机碳含量、储集层厚度、储集层孔隙度、储集层渗透率、盖层厚度、局部构造特征等总共 7 项与油气形成有关的地质数据，它们都是基础地质信息。其中的生油岩厚度与生油岩有机碳含量两项基础地质信息同属生油条件，因此生油条件就是由这两项基础地质信息构成的组合地质信息。同样，储集层厚度、孔隙度、渗透率三项基础地质信息构成了储油条件这个组合地质信息。归类后可以得到有层次关系的数据体系，见表 15-12。

表 15-12　地质数据体系表

组合地质信息	基础地质信息
生油条件	生油岩厚度、生油岩有机碳含量
储油条件	储集层厚度、孔隙度、渗透率
盖层条件	盖层厚度
构造条件	局部构造特征（例如构造面积、闭合度等）

(2) 生成基础地质信息图件

如果某些基础地质信息已有现成的图件并且符合要求，则不必再由基础地质数据去生成图件。但是，在多数情况下，由于图件比例尺不同，以及图件的收缩、变形、破损等原因，往往需要重新生成图件，即便是直接将原有图输入计算机，也要作位置校正及误差校正，否则叠合后

的地质信息将可能失真。

由基础地质数据生成平面图件就是用约定的插值计算方法，由计算机绘制等值线图或分带图。

(3) 生成组合地质信息叠合图件

由同类的基础地质信息图件，按约定的算法可以生成组合地质信息图件。例如由生油岩有机碳含量与生油岩厚度的等值线图或分带图，叠合后可以生成生油条件等值线图或分带图。

在叠合前，可以按每种基础地质信息在组合地质信息中所起作用的大小(重要性)，赋以相应的权系数，使各个基础地质信息起到不同的作用，从而使生成的组合地质信息图更为合理。

(4) 生成综合地质信息的二次叠合图

由若干个组合地质信息图件，按约定的叠合方法可以生成最后的综合地质信息图件。由基础地质信息图件到形成综合地质信息图件，总共经过了两次叠合过程。

同样，进行二次叠合前，仍然需要根据每个组合地质信息对探区油气形成所起作用的大小，赋以适当的权系数，以保证每种组合地质信息对最终评价起到应有的作用。

3. 地质数据的平面插值

在地质数据中，除地球物理、遥感等少数地质数据是密集采样外，其他绝大多数的地质数据都是稀疏的离散点值。因此，为了进行地质信息间的叠合，事先需要根据少数的离散点值扩充为平面密集的数值点。这一问题在数学上就是二维(平面)数据的插值问题。

这里介绍一种简单而适用的平面插值方法，即距离倒数平方加权法。如果探区中有 m 种可作为指导找油的基础地质数据，其中第 j 种基础地质数据经过挑选后可供使用的原始数据有 n 个，其第 i 个数据的平面坐标为(x_i, y_i)，观测值为 z_i。设平面插值域上任意一点 p 的平面坐标为(x, y)，p 点与第 i 个原始数据点的距离为 $D_i(i=1,2,\cdots,n)$，则：

$$D_i = [(x - x_i)^2 + (y - y_i)^2]^{\frac{1}{2}} \qquad (i = 1,2,\cdots,n) \tag{15-44}$$

p 点的插值 z 可用下面的公式求得：

$$z = \begin{cases} \sum_{i=1}^{n} z_i(D_i)^{-2} \Big/ \sum_{i=1}^{n} (D_i)^{-2} & (\text{所有的 } D_i \neq 0 \text{ 时}) \\ z_i & (\text{如果有 } D_i = 0 \text{ 时}) \end{cases} \tag{15-45}$$

由式(15-45)可以看出，平面上任意一点 p 的插值方式有两种可能：当 p 点与第 i 个原始数据点重合，即如果 $D_i=0$ 时，则 z 值等于 z_i。当 p 点与 n 个原始数据点都不重合，即如果所有的 $D_i \neq 0$ 时，则插值 z 受 n 个原始数据 z_i 的影响，而每个原始数据对 z 的影响程度，与原始数据到 p 点的距离平方成反比。可见，z 值的大小主要受靠近 p 点的原始数据影响，而远离 p 点的原始数据对 z 的影响较小。

需要说明的是，上面的插值法仅仅是最简单的一种方法。根据实际需要，计算时也可以约定其他插值方法，例如高次趋势面逼近法、克里金法、埃米尔插值法等。

由于各个探区的地质情况千差万别，即使是同一个探区中，也有若干个在性质上不同的次一级地质单元，因此，为了使插值结果符合实际地质情况，在插值时要对原始数据点的选择及使用数量作一些人为限定。一般归纳为以下几种方法。

(1) 全点插值法

全点插值法是对原始数据不加任何限定，即所有数据点都参加插值计算。当原始数据点比较少，而勘探人员对探区的地质结构又不太清楚时，最好用全点插值法。用这种插值法得到的图形比较光滑，不容易发生图形畸变。

(2) 近点插值法

近点插值法是从 n 个原始数据点中选用距离 p 点最近的 t 个点参加插值计算。当探区的地质结构比较复杂时，为了使参加插值计算的数据点限定在同一地质单元内，最好使用近点插值法。近点的个数 t，可视具体的地质情况确定。

使用近点插值法时，必须首先计算插值点到所有原始数据点之间的距离，并且按距离的远近，从小到大依次排序，而后从中选出 t 个与 p 点距离最近的原始数据点。

(3) 圆内插值法

圆内插值法是以插值点 p 为圆心，选择合适的半径 R 划圆，只使用圆周内点的原始数据参加插值计算。

(4) 象限插值法

象限插值是以插值点 p 为中心，将插值域分为若干个象限，一般可分为 4 个象限或 8 个象限，在每个象限中选择距 p 点最近的 t 个原始数据点参加计算。这种插值方法可以保证所使用的数据点在各象限方向上的均匀性，以消除由于数据点分布不均匀造成的方向性干扰。

需要指出，对原始数据点的使用数量以及选点方法都是人为确定的，所以绝非仅有以上 4 种方法，而应当按实际需要灵活确定。

4. *多种信息叠合方法*

各种基础地质数据经过插值计算后，进一步可以得到等值线图或分带图。如果按已形成的数据体系，把同一类的基础地质信息图件重叠在一起，按某种约定的算法进行叠合，则可以得到组合地质信息图。再经过二次叠合就能得到最终的预测探区含油气有利地带的多种信息叠合评价图。

(1) 叠合前的数据预处理

① 地质数据的正规化。进行信息叠合前，每种基础地质数据都要经过标准化处理，使它们变换到同一尺度范围内，以保证各种地质信息之间的等价性。一般使用极差正规化方法，将各种原始数据变换到[0,1]区间范围内。

② 对地质信息赋权值。进行信息叠合之前，应当根据每种基础地质信息或组合地质对油气形成所起的作用，赋以合理的权系数，以体现叠合时它们所起的作用大小。

(2) 一次叠合方法

一次叠合方法是指由基础地质信息叠合生成组合地质信息的方法。

① 乘积叠合。这种叠合方法是把平面上同一坐标点的 m 种基础地质信息值进行连乘，得到该点的组合地质信息 z_j，即：

$$z_j = \prod_{i=1}^{m} z_{ji} \qquad (j = 1,2,\cdots,w) \tag{15-46}$$

式中 z_j——经过一次叠合生成的第 j 种组合地质信息；

z_{ji}——第 j 种组合地质信息中的第 i 种基础地质信息。

w——组合地质信息种类个数。

m——第 j 种组合地质信息所含的基础地质信息种类数。

用乘积叠合图对探区进行分带评价时，其分带区间(等值线距)是不等间距的，区间间隔值可按如下公式计算：

$$H_r = \left[\frac{1}{k}(r-1)\right]^m \qquad (r = 1,2,\cdots,k+1) \tag{15-47}$$

式中　H_r——第 r 个分带区间的间隔值；

k——分带区间总数；

m——进行叠合的基础地质信息个数。

② 累加叠合。这种叠合方法是把平面上同一坐标点的 m 种基础地质信息值进行累加，得到组合地质信息 z_j，即：

$$z_j = \sum_{i=1}^{m} z_{ji} \qquad (j = 1,2,\cdots,w) \tag{15-48}$$

对于累加叠合，其分带区间是等间距的，区间间隔值的计算公式如下：

$$H_r = \left[\frac{1}{k}(r-1)\right]m \qquad (r = 1,2,\cdots,k+1) \tag{15-49}$$

③ 集合取小叠合。这种叠合方法是把平面上同一坐标的每种基础地质信息作为一个元素，如果有 m 种基础地质信息，则可构成一个由 m 个元素组成的一个集合。叠合时是从集合中取出数值最小的元素作为叠合值，即：

$$z_j = \min_{1\leqslant i\leqslant m}(z_{ji}) \qquad (j = 1,2,\cdots,w) \tag{15-50}$$

对于集合取小叠合，其分带区间也是等间距的，区间间隔值的计算公式如下：

$$H_r = \frac{1}{k}(r-1) \qquad (r = 1,2,\cdots,k+1) \tag{15-51}$$

(3) 二次叠合方法

二次叠合是指经过一次叠合生成的 w 种组合地质信息，再进行二次叠合，生成综合地质信息。如果归类后的数据体系有两个层次，那么，由基础地质信息到最后的综合地质信息，按上述三种叠合方法的组合，总共有 3×3=9 种叠合方法。

① 双重乘积叠合

这种叠合是指由基础地质信息经过乘积叠合生成组合地质信息，再由组合地质信息经过乘积叠合生成综合地质信息。计算公式为：

$$z = \prod_{j=1}^{w} z_j = \prod_{j=1}^{w}\prod_{i=1}^{m} z_{ji} \tag{15-52}$$

式中　z——综合地质信息，直接由基础地质信息求得；

z_j——第 j 种组合地质信息；

z_{ji}——第 j 种组合地质信息中的第 i 种基础地质信息。

② 乘积累加叠合

$$z = \prod_{j=1}^{w} z_j = \prod_{j=1}^{w}\sum_{i=1}^{m} z_{ji} \tag{15-53}$$

③ 乘积取小叠合

$$z = \prod_{j=1}^{w} z_j = \prod_{j=1}^{w}(\min_{1\leqslant i\leqslant m} z_{ji}) \tag{15-54}$$

④ 双重累加叠

$$z = \sum_{j=1}^{w} z_j = \sum_{j=1}^{w}\sum_{i=1}^{m} z_{ji} \tag{15-55}$$

⑤ 累加乘积叠合

$$z = \sum_{j=1}^{w} z_j = \sum_{j=1}^{w}\prod_{i=1}^{m} z_{ji} \tag{15-56}$$

⑥ 累加取小叠合

$$z = \sum_{j=1}^{w} z_j = \sum_{j=1}^{w} (\min_{1 \leqslant i \leqslant m} z_{ji}) \tag{15-57}$$

⑦ 双重取小叠合

$$z = \min_{1 \leqslant j \leqslant w} (\min_{1 \leqslant i \leqslant m} z_{ji}) \tag{15-58}$$

⑧ 取小乘积叠合

$$z = \min_{1 \leqslant j \leqslant w} \left(\prod_{i=1}^{m} z_{ji} \right) \tag{15-59}$$

⑨ 取小累加叠合

$$z = \min_{1 \leqslant j \leqslant w} \left(\sum_{i=1}^{m} z_{ji} \right) \tag{15-60}$$

(4) 各种叠合方法的适用性

乘积、累加、集合取小叠合方法的地质含义是不相同的。

乘积叠合方法，适用于被叠合的各种地质信息的乘积值大体上可以代表一个新的地质变量或者具有某种特定的地质含义的情况。例如，生油岩的厚度与氯仿沥青含量的乘积值，是与生油潜量有关的组合地质信息，大体上相当于生油丰度。

累加叠合方法，适用于各种被叠合的地质信息之间的关系尚不清楚的情况，而经累加叠合后的组合地质信息表示各种与油气形成有关的地质信息的总和。累加叠合后的数值越大说明含油气的可能性越大。

集合取小叠合方法是出于最小因素思想，例如在一个勘探地区，如果生油、储油、盖层等控制油气形成的必要地质条件中，只要其中有一项不具备油气形成条件，则形成不了油气藏。可见，集合取小叠合是从最保险的角度出发的。

当然，使用哪种叠合方法进行一次叠合和二次叠合，以及在叠合前对基础地质信息或组合地质信息所赋的权系数的大小，都要根据具体地质条件，由熟悉探区情况的有关人员讨论商定。

5. 算例

我国北方某沉积盆地是一个中新生代断陷盆地，沉积岩厚度达到 5000 m 以上，具备形成油气的基本地质条件，有条件建成一个新的产油区。盆地中的 M 坳陷是最有远景的含油地区，因而评价该坳陷内各个地带的含油气地质条件，并在坳陷内寻找最有利的勘探地带是十分重要的。目前，勘探的主要目的层系是 k 系 p 组，钻探工作主要集中在坳陷的东部地区。全坳陷已基本完成地震勘探工作，因此用多种信息叠合法进行评价时，是以地震勘探资料为主，再加上已有的钻井资料，对全坳陷进行有利勘探地带的预测。

根据地震资料及钻井资料，选用生油岩厚度、*TTI* 值、生油岩沉积相、储集层厚度、储集层沉积相、局部构造面积、幅度、类型、钻井油气显示、盖层厚度等共 10 项基础地质信息。由这 10 项基础地质信息经过一次累加叠合后形成生油条件、储油条件、构造条件、含油气状况、盖层条件总共 5 项与油气形成有关的组合地质信息。

根据地震、钻井资料，在 M 坳陷已发现了 76 个局部构造。所选用的 10 种基础地质信息，按统一标准划分为一、二、三总共 3 个级别，分级标准与地质信息间的关系见表 15-13。表中的 3 个级别中三级的含油气条件最好，二级居中，一级最差。

表 15-13　M 坳陷的地质信息及分级标准表

组合地质信息	基础地质信息	一　级	二　级	三　级
生油条件	生油岩厚度/m	<300	300～500	>500
	TTI 值	<8	>256	8～256
	生油岩沉积相	其他相	浅湖相	湖　相
储油条件	储集层厚度/m	<100	100～300	>300
	储集层沉积相	湖沼相	河流平原相	冲积扇三角洲相
构造条件	构造面积/km^2	<20	20～35	>35
	构造幅度/m	<100	100～150	>150
	构造类型	断块、岩性	断鼻、半背斜、潜山	背　斜
含油气状况	油气产状	干　井	油气显示	油气流
盖层条件	盖层厚度/m	<1200	>2400	1200～2400

为了计算方便，将基础地质信息按一、二、三级分别赋值为 1、2、3。按分级标准，形成 M 坳陷 76 个局部构造的原始数据。原始数据的坐标均以原图的左下角为坐标原点(0,0)，各点的纵、横坐标值是以 cm 为单位在原图上实际测量得到。

5 种组合地质信息经二次累加叠合生成最终的多种信息叠合评价图。一次以及二次叠合过程见图 15-11。

叠合计算时选用的计算参数见表 15-14。按这些选定的参数，经过计算得到 10 张基础地

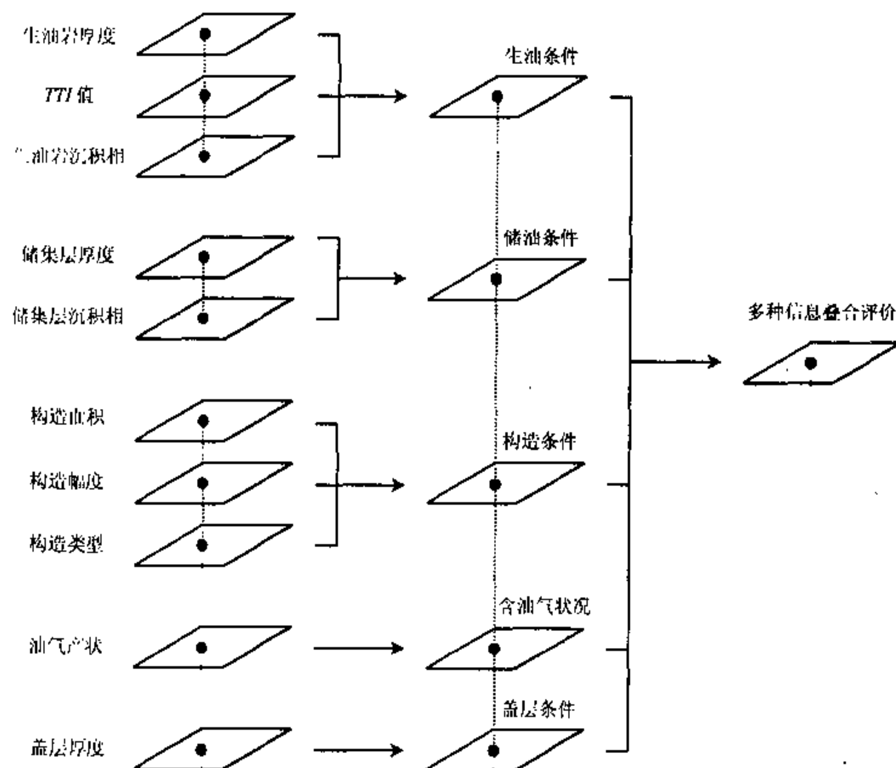


图 15-11　地质信息叠合过程示意图

质信息的平面插值图。按组合方式由这10张图经过一次累加叠合生成5张组合地质信息图，即：生油条件评价图、储油条件评价图、构造条件评价图、含油气状况图、盖层条件评价图。由这5张图经过二次累加叠合生成多种信息叠合评价图(综合地质信息图)，如图15-12所示。

表15-14　叠合计算参数表

基础地质信息名称	基础地质信息个数	平面插值方法	一次叠合时的权系数	一次叠合方法	组合地质信息名称	组合地质信息中的基础地质信息个数	二次叠合时的权系数	二次叠合方法
生油岩厚度	76	圆内法	1.0	累加叠合	生油条件	5	1.0	累加叠合
TTI 值	76	圆内法	1.0					
生油岩沉积相	76	圆内法	1.0					
储集层厚度	76	圆内法	1.0	累加叠合	储油条件	2	1.0	
储集层沉积相	76	圆内法	1.0					
构造面积	76	圆内法	1.0	累加叠合	构造条件	3	1.0	
构造幅度	76	圆内法	1.0					
构造类型	76	圆内法	1.0					
油气产状	76	圆内法	1.0	累加叠合	含油气状况	1	1.0	
盖层厚度	76	圆内法	1.0	累加叠合	盖层条件	1	1.0	

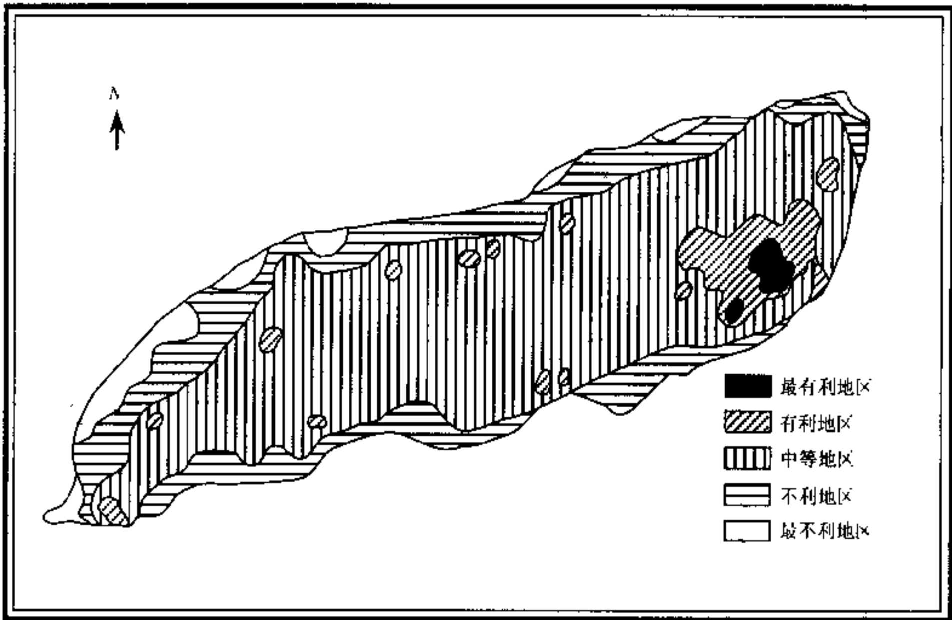


图15-12　多种信息叠合评价图

这些评价图统一划分为5个级别的分带，即：含油气最有利地带、有利地带、中等地带、不利地带、最不利地带。

根据计算结果，选出了21个含油气地质条件较好的局部构造，其中的4个局部构造经钻探已证实为含油构造。后期实践验证，图15-12中东部的最有利与有利地区，建成了具有一定规模油气产能的油田。

附录　SURFER 环境下部分绘图基本子程序

初始化子程序 in

```
      Subroutine in character * 8 fn
      write( * ,'(a ))')' Enter Plot filename:'
      read( * ,'(a)') fn
      if(fn. eq. 'q'. or. fn. eq. 'Q') Stop 'Stop ! '
      open(19,file=fn,StatuS='unknown')
      write(19,3) 2
3     format('FL',1x,i2)
      write(19,1) 0,0
1     format('PI',2(1x,i1))
      write(19,2) 0,0
2     format('TR',2(1x,i1))
      return
      end
```

绘图比例子程序 fact

```
      subroutine fact(a)
      write(19,1) a * 0. 1,a * 0. 1
1     format('SC',2(1x,f6. 3))
      return
      end
```

选绘图笔子程序 pen

```
      subroutine pen(i)
      write(19,1) i
1     format('SP',1x,i2)
      return
      end
```

绝对抬笔移动子程序 movea

```
      Subroutine movea(x,y)
      write(19,1) x,y
1     format('MA',2(1x,f10. 3))
      return
      end
```

绝对落笔画线子程序 linea

```
      Subroutine linea(x,y)
      write(19,1) x,y
```

```
1     format('PA',2(1x,f10.3))
      return
      end
```

相对抬笔移动子程序 mover

```
      subroutine mover(x,y)
      write(19,1) x,y
1     format('MR',2(f8.2))
      return
      end
```

相对落笔画线子程序 liner

```
      subroutine liner(x,y)
      write(19,1) x,y
1     format('PR',2(1x,f8.2))
      return
      end
```

绘实型数子程序

```
      Subroutine numb1(x,y,h,a,f)
      write(19,1) x,y,h,a,f
1     format('PS',2(1x,f10.3),2(1x,f10.3),1x,'"',f6.2,'"')
      return
      end
```

绘整型数子程序

```
      Subroutine numb2(x,y,h,a,i)
      if(i.ge.0) then
      if(i.le.9) then
      write(19,1) x,y,h,a,i
      else
      if(i.le.99) then
      write(19,2) x,y,h,a,i
      else
      if(i.le.999)then
      write(19,3) x,y,h,a,i
      else
      write(19,4) x,y,h,a,i
      end if
      end if
      end if
      else
      if(i.ge.-9) then
      write(19,2) x,y,h,a,i
```

```
      else
      if(i.ge. -99) then
      write(19,3) x,y,h,a,i
      else
      if(i.ge. -999)then
      write(19,4) x,y,h,a,i
      else
      write(19,5) x,y,h,a,i
      end if
      end if
      end if
      end if
1     format('PS',2(1x,f10.2),2(1x,f10.2),1x,'"',i1,'"')
2     format('PS',2(1x,f10.2),2(1x,f10.2),1x,'"',i2,'"')
3     format('PS',2(1x,f10.2),2(1x,f10.2),1x,'"',i3,'"')
4     format('PS',2(1x,f10.2),2(1x,f10.2),1x,'"',i4,'"')
5     format('PS',2(1x,f10.2),2(1x,f10.2),1x,'"',i5,'"')
      return
      end
```

选择符号库子程序

```
      subroutine setsty(asc)
      character *(*) asc
      write(19,1) asc
1     format('SS',1x,'"',a,'"')
      return
      end
```

绘字符串子程序

```
      Subroutine text(x,y,h,a,asc)
      character *(*) asc
      write(19,1) x,y,h,a,asc
1     format('PS',2(1x,f12.2),2(1x,f10.2),1x,'"',a35,'"')
      return
      end
```

参 考 文 献

1 于志钧，赵旭东. 石油数学地质. 石油工业出版社，1987

2 赵旭东. 石油数学地质概论. 石油工业出版社，1992

3 陆明德，田时芸. 石油天然气数学地质. 中国地质大学出版社，1991

4 王学仁. 地质数据的多变量统计分析. 科学出版社，1982

5 Harbaugh J. W 等著. 罗人彦，何保侃等译. 地质过程的计算机模拟. 地质出版社，1986

6 油气资源评价方法研究与应用编委会编. 油气资源评价方法研究与应用. 石油工业出版社，1988

7 石广仁. 油气盆地数值模拟方法. 石油工业出版社，1994

8 李汉林，赵永军. 计算机绘制地质图. 石油大学出版社，1997

9 刘承祚，孙惠文. 数学地质基本方法及应用. 地质出版社，1982

10 余金生、李裕伟. 地质因子分析. 地质出版社，1985

11 中山大学数学力学系. 概率论及数理统计. 人民教育出版社，1980

12 李汉林，赵永军. 岩性识别的多元统计方法. 见《地质论评》，V. 1，No. 1，1998

13 赵永军，李汉林. 油气地表化探失真数据的概率判定与处理. 见《地质论评》，V. l. 40，Sup. 1994

14 Yukler，M. A. ，C. Cornford，and D. H. Welte，1978，One－dimensional model to simulate geologic，hydrodynamci and thermodynamic development of a sedimentary Basin：Geol. Rundschau，V. 67，P. 960－979.

15 Welte，D. H. ，and M. A. Yukler，1981，Petroleum origin and accumulation in basin evolution－A quantitative model：AAPG Bulletin，V. 65，P. 1378～1396.

16 Tissot，B. P. ，and D. H. Welte，1984，Petroleum formation and occurrence，2d ed：New York，Springer－Verlag.

17 Nakayama，K. ，and D. C. V. Siclen，1981，Simulation model for petroleum exploration：AAPG Bulletin，V. 65，P. 1230－1255.

18 Nakayama，K. ，1978，Hydrocarbon－expulsion model and its application to nigata Area，Japan：AAPG Bulletin，V. 71，p. 810－821.

19 Waples，D. W. ，1980，Time and temperature in petroleum formation：Application of Lopatin's method to petorleum exploration：AAPG Bulletin，V. 64，P. 916～926.

20 Tissot，B. P. ，R. Pelet，and P. Ungerer，1987，Thermal history of sedimentary basins，maturation indices，and kinetics of oil and gas generation：AAPG Bulletin，V. 71，P. 1445～1466.

21 Magara，K. ，1978，Compaction and fluid migration－Practical petroleum geology：Amsterdam－Oxford－New York，Elsevier Scientific Publishing Company.

